

Use of Biotechnology in Nutrition Research

Christopher M. Ashwell
Department of Poultry Science
North Carolina State University
Campus Box 7608, 134E Scott Hall
Raleigh, NC 27695

Phone: 919-513-7335, Fax: 919-515-2625
chris_ashwell@ncsu.edu

Classic animal nutrition research has involved the determination of nutritional requirements, diet formulation, and the monitoring of animal performance. Historically, specific nutrition research often involved the assessment of targeted pathways such as carbohydrate or lipid metabolism where biochemical and enzyme assays monitored the effects of diet. More recently with the development of the field of molecular biology, researchers have been able to study the impact of diet on the organism at the molecular level. The focus of this review is to summarize new research frontiers and to highlight the information biotechnology can provide to nutrition research.

The term “nutrigenomics” was coined in 2002 to describe the trend in human nutrition research towards individualized dietary formulation based on how diet effected the expression of specific genes and the development of specific diseases including diabetes, obesity, and heart disease. In reality the scope of this field is significantly larger.

The interaction of an organism with its diet (or nutrition source) is an intimate and complex physiologic affair that is based on multiple organ systems working in concert. The regulation of these processes can be found at all levels from genetics, gene expression, proteins, to specific metabolites. Only until recently has the technology become available to follow the regulation of these processes. Nutrition researchers are just beginning to utilize tools to ask scientific questions about diet involving genomics, functional genomics (gene expression), proteomics, and matabolomics. This review will address the capabilities and limitations of these technologies and the potential impact of a systems-wide approach on the development of nutritional science.

Introduction

Dietary components (nutrients) were long thought of as a source of energy or as cofactors until the discovery of metabolites role in the regulation of enzymatic activity through allosteric control. Metabolites were also shown to regulate the secretion of hormones. Later these effects were found to be the result of modified gene expression. Biotechnology has provided evidence of two key concepts involving the interaction of diet (nutrition) and the individual organism. First, nutrition affects the individual through the modulation of expression of genetic information in response to diet composition. Second, nutritional effects must be characterized on the basis of the genetics of the individual organism.

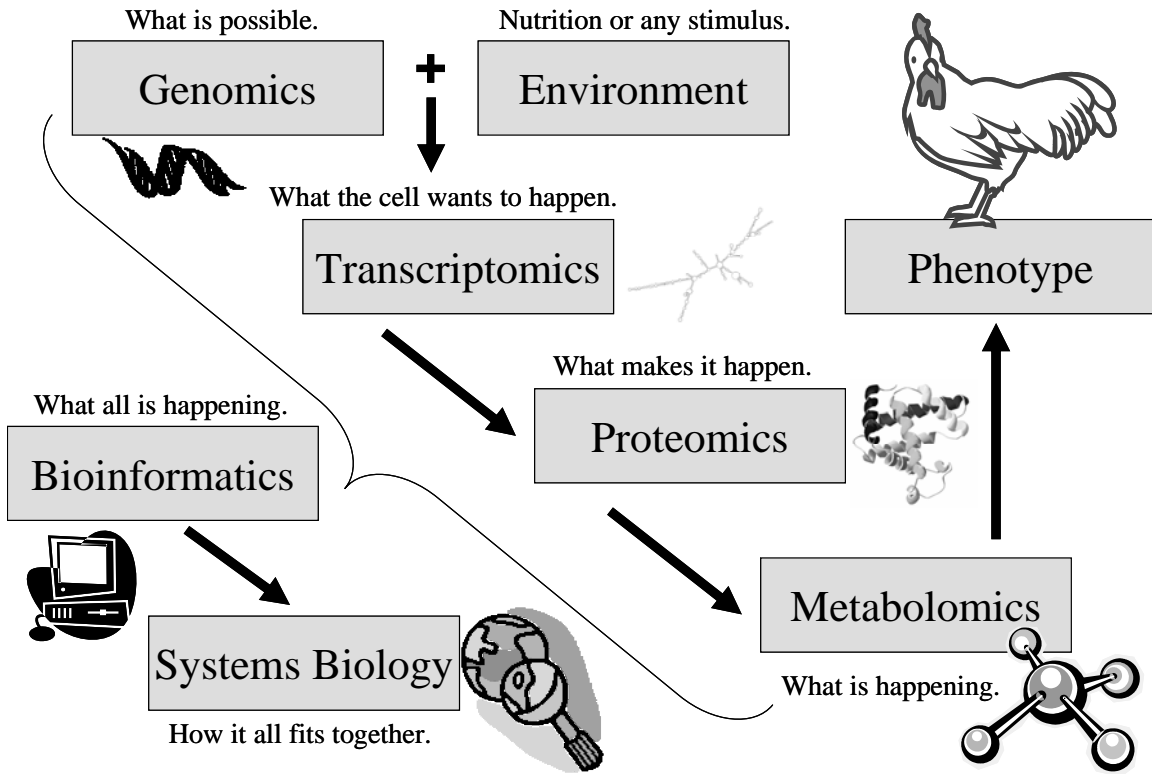
Adaptation to dietary components by the organism implies the regulation of physiologic and metabolic pathways. Most of our knowledge on the effects of nutrients on gene expression has been acquired in animal models, many examples of which can be found in a recent overview of the mechanisms by which nutrients interact with their molecular targets to modify gene expression (Muller and Kersten, 2003).

Variation of the individual response to diet can be explained by the underlying differences in genetics across a population. This observation holds true in human as well as most outbred animal populations. This variation or polymorphism is the basis for individuality and results from sequence differences within the genome of an organism. The effects of each varying sequence are minute but the culmination of the differences across the entire genome produces an individual's genotype or genetic potential. This genotype interacts with environmental factors to produce the phenotype or outward appearance or performance of an individual. Numerous studies in animals and humans have shown that individual genotypic variations can alter nutrient metabolism, from relatively mild conditions like lactase gene polymorphisms that result in lactose intolerance to potentially severe pathological conditions like phenylketonuria (Harvey et al 1998).

It is clear that the quantity and quality of the diet modulates the expression of numerous genes in various tissues. Each individual will respond to a specific diet in a unique manner concomitant with their genetic profile of polymorphisms. One of the goals of nutrigenomics research is the development of consensus responses to specific dietary stimuli so that anomalies can be identified and studied further. It is these anomalies that will provide the basis for understanding how genetic differences are associated with specific response to nutrients, how genetic differences in combination with diet result in maladies like obesity, and how specific nutrient have beneficial or deleterious effects.

Research using the knowledge from the Human Genome Project will ultimately enable scientists to understand the functions of human genes and how they are regulated. This knowledge will provide them with information on how genes and nutrients interact and the effect of individual genetic differences on diet and nutrition. This research will be directly applicable to other species whose genome sequencing projects are underway including many livestock species. Research can help to identify these effects and help to understand why certain nutrients and foods are of benefit to health. Developing within this genome era were technologies that were increasingly broad in scope that included automation, high throughput, and data intensive. Many of these technologies also involved miniaturization of standard techniques to suit the new high throughput experimental designs. A description of the state of these technologies and their implications on nutrition research follows including the study of genomics (polymorphism), functional genomics (gene expression), proteomics (protein expression), metabolomics (metabolite profiles), bioinformatics (data storage and mining), and systems biology (integrated data analysis). The organization of these processes and data sources is found in Figure 1.

Figure 1: Diagram of the integration of molecular technologies and the flow of genetic information from the genome by its interaction with the environment through gene expression (transcription), protein expression (translation), and metabolites to the ultimate phenotype. The accumulation of information (bioinformatics) and its explanation (systems biology) are also included.



Genomics

Genomics concerns the analysis of DNA, the genome, and focuses on identifying the variation in the genome between individuals. Genomics tries to correlate this variation with phenotypic parameters (linkage and association studies). Variation in the sequence at the nucleotide level (substitution and small deletions, duplications, insertions) has long been considered most important; variant coding sequences directly translate to variant function and variation can be used for linkage and association studies. New quantitative technologies including array-CGH (comparative genomic hybridization) and quantitative SNP technology revealed an unexpected and large-scale copy number variation (CNV, deletions and duplications) in the human genome (Iafra et al 2004, Sebat et al 2004). The majority of CNVs include genes, thereby directly influencing the expression level (in theory 50% up or down), and should be considered as an important and previously neglected type of variation in the human genome. For measurement, SNPs (single nucleotide polymorphisms) are currently the most popular tools, although di-nucleotide repeats (tandem repeats of two nucleotides such as CA or AT) and AFLP (amplified fragment length polymorphisms, mainly employed in plants) are also widely applied. Variation and recombination in the genome is not spaced randomly and efforts currently are attempting to characterize the nature of these “hot-spots”. Currently high-throughput

genomewide analysis is facilitated using several technologies, the most powerful being array technology (DNA chips and micro-arrays), mass spectrometry (e.g. MALDI-TOF) and beads-based flow sorting. The maximal capacity lies in the order of 100,000 SNP typings per day. The current major problem is not technological but financial, with the present cost per genotype of \$0.01-0.05.

Gene Expression- Transcriptomics

Focusing on the analysis of RNA (the transcriptome), transcriptomics aims at measuring the level of expression of all or a selected subset of genes based on the amount of RNA present in a sample. Currently, the most powerful tool available is DNA array technology. Using one array the expression level of up to 50,000 transcripts can be measured in parallel, and tens of samples can be screened per day. In these studies, hundreds of genes are usually varying in expression. The difficulty is to organize the results in such a way that they can be used to elucidate biological mechanisms, or to derive biological markers for a given physiological situation. Such data treatment is obviously an essential requirement if one wants to understand the overall consequences of nutrient intake.

The main limitation lies in the sensitivity of the assay as well as in data analysis. Statistically significant measurements can sometimes only be obtained for the most abundantly expressed genes, and when expression differences are changed by a factor of two or more. When smaller changes need to be detected, the measurement has to be repeated several times, making studies rather costly.

As for any new technology, array-based transcriptomics is hindered by initial limitations in analytical precision and standardization. The standardization issue was noted at an early stage. The Microarray Gene Expression Database (MGED) group was founded in 1999 with the goal of facilitating the adoption of standards for DNA array experiment annotation and data representation, as well as the introduction of standard experimental controls and data normalization methods (www.mged.org). In addition, high per-analysis cost seriously reduces the number of measurements performed per study. Furthermore, different platforms are used, e.g. cDNA vs. oligonucleotide array and printed microarrays vs. on chip synthesis, and designs are regularly modified to incorporate new genes and improved probe sequences, thereby complicating data comparison (Nimgaonkar et al 2003). Several commercial suppliers produce off-the-shelf arrays (e.g. Affymetrix, Agilent, Amersham) or oligonucleotide collections for custom spotting (e.g. Illumina, Operon). Intrinsically, these provide a first 'standardization' which is desperately needed to be able to compare results from different studies. Because of these problems, studies that adequately meet rigid statistical requirement are in fact relatively scarce (Kothapalli et al 2003). In view of these weaknesses, data from array-based transcriptomics need to be interpreted cautiously.

Finally, it should not be underestimated that the major source of variability arises from the starting biological material itself. Ensuring the validity of this material, collecting all possible variables (from genome, to age and environment) and controlling the sampling conditions and timing are essential to obtain meaningful data. In this case, the use of

inbred lines is valuable in controlling the genetic variation. Another limitation lies in the tools used for data analysis (software). Current bioinformatic tools are somewhat effective but opinions vary considerably regarding the best computational algorithms to apply.

Proteomics

Proteomics technology focuses on the analysis of proteins and their interactions. The challenge lies in the development of technologies, which are able to cope with the huge differences in chemical properties of proteins as well as the wide dynamic range of protein concentrations. Initially, two-dimensional (2D) gel electrophoresis was used to measure the expression level of a large number of proteins. If the full set of proteins separated by 2D gels is to be identified, automated equipment is used for excising protein spots, digesting the proteins therein, and analyzing the resulting peptides using mass spectrometry. However, 2D-gel analysis is biased towards the most abundant changes, which might lead to erroneous conclusions since also subtle variations may lead to important changes in metabolic pathways. In addition, low abundant proteins and very hydrophobic, acidic or basic proteins are often not detected and identification of the proteins resolved is time consuming and costly (Gygi et al 2000). Like with array technology, 2D-gel analysis can be combined with two-color fluorescent labeling, highlighting those proteins that differ in expression between the two samples.

Recently, mass spectrometry (MS) has come into play as an exciting and very powerful analytical tool. Proteins are submitted to proteolytic degradation to form collections of peptides that are subsequently analyzed using MS (peptide mass fingerprints). When the molecular mass of certain peptides does not concur with published structures, these peptides are further characterized by MS-MS tandem mass spectroscopy. Such analyses will also reveal post-translational modifications of proteins, such as phosphorylation. MS can be used to quickly determine the identity of a specific protein and it facilitates analysis of very complex protein samples, zooming in on those proteins that differ in expression. Impressive studies have been performed in the area of cancer research and diagnosis using mass spectrometry in combination with 2D liquid chromatography (2D-LC). The limiting step for this technology lies mainly in data analysis, i.e. computing power and the lack of adequate software tools. Generating MS traces is a matter of seconds with cost of the measurement being a minor issue. Another approach uses isotope coding and subsequent quantification relying on digestion of the protein mixture and separation on peptide rather than on protein level (shotgun proteomics). While gel-associated drawbacks are circumvented, isotope coding is limited by the risk of insufficient yields and the alteration of the sample composition. Moreover, it is often compromised by a protein bias due to tagging post-digestion, amino acid-targeted reagents and possible chromatographic separation of the light and heavy labels.

Major challenges for the field of proteomics are posed by the study of protein-protein interactions and the relationship between protein polymorphisms – nutritional value. To study protein-protein interactions, two-hybrid cloning systems have proven to be efficient and successful techniques (Fields and Song 1998). Recently, two-hybrid arrays have been developed, in which the screening is performed in a colony array format with each colony

expressing a different pair of proteins (Cagney et al 2000). Array screens can be easily automated facilitating high-throughput and reproducible protein-protein interaction screens. Furthermore, comparing the results from several assays circumvents the problem that single assays generate a high number of false positives.

Most biological questions require the differential analysis of two biological states, typically case vs. control, and, consequently, quantitative tools. It should be noted that while quantitative transcriptomics have evolved into commercialized and partially standardized platforms, quantitative proteomics is just emerging and standards are largely unavailable.

Proteins in body fluids like milk are translated from mRNAs that are expressed in different tissues, such as mammary epithelial cells and milk leukocytes. Consequently, milk proteins cannot be readily deduced from transcriptomics, and proteomics has particular relevance here. Applications include identification of minor milk proteins with potent biological function, such as growth factors, and investigation of milk protein polymorphisms. Such polymorphisms have implications for the properties and processing of milk, as well as for its nutritional value, which is determined not only by amino acid composition but also by digestibility and digestion rate of proteins (Fitzgerald et al 2003). Moreover, polymorphism may change the pattern of peptides released during digestion in the gastrointestinal tract, which may result in differences in biological activity and allergenicity of their peptide mixes.

Metabolomics

Metabolomics technology focuses on the analysis of metabolites, the metabolome. It tries to measure the level of all substances (other than DNA, RNA or protein) present in a sample; the metabolome comprises the complete set of metabolites synthesized by a biological system. Such a system can be defined by level of biological organization, such as organism, organ, tissue, cell, or cell compartment levels. Today, the best tools for metabolomics research are proton nuclear magnetic resonance (NMR) and mass spectrometry (MS). Biologically relevant samples can easily be obtained from blood, sweat, urine, and feces.

Metabolomic analyses have only just begun to be in wide use. NMR-based metabolite profiling through highly quantitative and broad-spectrum classes still suffers from inherent sensitivity issues. The dynamic range is between few nanomoles to few hundred micromoles. Thus most abundant metabolites, i.e. steady-state concentrations are easily observed. However, recent cryoprobe technology shows great promise to overcome sensitivity hurdles. LC-MS, complementary to NMR, offers superb sensitivity but is limited by the essentially nonquantitative nature of mass spectrometry requiring internal standardization. Addressing automatic spectra processing, a few obstacles are worth mentioning: although peak-recognition software is getting smarter (for NMR and MS), inconsistencies in chemical shifts and baseline shifts in NMR spectra have to be compensated for and this is not a standardized task.

Publicly available metabolomic databases are being developed similarly to genomic sequence repositories and technical advances as well as improved data mining and

analysis tools are in development. Due to pleiotropic effects, the effect of a nutrient may lead to changes of metabolite levels in various, seemingly unrelated biochemical pathways. Therefore, a comprehensive analysis of all metabolites is required to understand such hidden relationships. Both sample preparation and data acquisition must aim at identifying all classes of compounds, assuring high recovery as well as experimental robustness and reproducibility. The novelty of these approaches is evident since the first meeting of the Metabolomics Society was convened in June of 2005.

Bioinformatics and Data Mining

Bioinformatics is the technology enabling the data processing, clustering, dynamics, integration and storage of the overwhelmingly complex data sets produced by modern molecular research. Bioinformatics will play a crucial role to condense the massive amounts of data generated through high-through-put experimental procedures and to integrate these with data obtained from traditional techniques. The challenge is to combine all pieces of information so that all data can be looked at in a coherent way. Novel algorithms, software and hardware are being developed to translate sets of gene, protein and metabolite data into biochemical pathways.

In order to gain full access to these emerging powerful tools, it is paramount to address the enormous challenge of unifying complex and dissimilar data (Brazma, 2001). The incorporation of observations from numerous sources and domains into a unified, seamlessly searchable database and turning in it into knowledge will impact every facet of modern nutrition (Desiere et al 2002). Two important features are required in order to integrate data between databases; they have to speak the same language, and use the same identifiers for the same object. In other words, the biological domain should be described using specific vocabularies and ontologies (www.ebi.ac.uk/GOA/project.html) to allow meaningful data comparisons. The gene ontology consortium has developed a dynamic, structured, and precisely defined vocabulary for describing the roles of genes and their products in any organism. The goal of this effort is to address the deficits of the current rather divergent nomenclature schemes (Ashburner et al 2003). One of the best known providers of a human genome gene index and gene annotation, the Ensembl project (www.ensembl.org) is an entirely Open Source project and has been widely adopted by academic and commercial organizations (Hubbard et al 2002).

Major challenges that remain to be addressed are to define ‘normal’ and ‘healthy’ versus ‘unhealthy’ profiles, especially in the pre-disease stage. Sampling location and timing will have to be optimized within ethical and practical constraints. In outbred populations like humans, detection of subtle effects may require numbers of participants that are too large to be realistic, unless participants can be pre-selected based on their genotype.

Systems Biology

By helping to understand the interaction between nutrients and molecules in our bodies, the implementation of molecular biology and biochemistry in ‘classical nutrition’ research, followed by the technological revolution of molecular technologies described above, will greatly affect nutritional sciences. The first studies that span the levels of genome, transcriptome, proteome, and metabolome demonstrate this impact. Most of

these studies investigate differential effects on the level of (metabolic) pathways, and provide new mechanistic insights. However, the real potential of technologies does not limit itself to such differential display type of strategies, where the measurement of a very large set of parameters is exploited only for those parameters that show dramatic differences. The complete dataset contains significantly more information.

On top of that, the various 'layers' of each technology platform are of course related (genes encode RNA, which encodes the enzymes that catalyze the conversion of metabolites). Thus, in combining the datasets of genome, transcriptome, proteome and metabolome, a wealth of added information becomes available. In fact, this combination of datasets paves the way to a complete description of the biological behavior of a cellular system, in response to external stimuli.

Although the complexity of this proposed integration (i.e. systems biology) is exceeding the current bioinformatics tools and capacities, its implications for nutritional research can be enormous. Unlike biomedical interventions (drug therapy), nutrition is chronic, constantly varying, and composed of a very large amount of known and unknown bioactive compounds. Furthermore, nutrition touches the core of metabolism by supplying the vast majority of ingredients (both macro- and micronutrients) for maintaining metabolic homeostasis. This homeostasis stretches from gene expression to metabolism and from signaling molecules to enzyme cofactors. Thus, nutrition by its nature *needs* to be studied in an integrated way. So far, most of the tools for this integration were lacking, thus maintaining an unbridgeable gap between classical nutrition (studying human physiology with excursions into biochemical pathways) and biomedical sciences (elucidation disease-related molecular mechanisms). In applying systems biology to nutritional science, these paradoxical extremes are bridged and the complexity of the relationship between nutrition and health can be met by the complexity of the integrated approach. At the moment this is little more than a dream, since only premature and pragmatic example studies of this concept are currently being performed. Many hurdles need to be taken, most of them in the field of bioinformatics, before this discipline matures.

In the relation between nutrition and health (unlike the relation between nutrition and disease) it is necessary to develop a new concept of biomarker. It needs to reflect subtle changes in homeostasis and the efforts of the body (cellular systems, organs, and inter-organ interactions) to maintain this homeostasis. Also, it preferably should include a wide variety of biological actions. Furthermore, both efficacy and safety aspects should be monitored simultaneously. Single nutrients may have multiple known and unknown biochemical targets and physiological actions, which may not be easily addressed with classical biomarkers (i.e. the 'single-gene, protein or metabolite' approach, usually at non-physiological conditions). In addition, the efficacy assessment of health effects of nutritional components is even further complicated by the fact that single dietary constituents are hardly consumed as separate entities but are part of a dietary mixture.

The inter-individual variation at the genetic level, as discussed above, appears to be adding even more complexity to the nutrigenomics picture. Quite a number of these genetic polymorphisms have been described from a clinical genetics point of view,

simply because they pre-dispose to a pathological condition. These usually are the monogenic forms, with a pronounced effect on functionality. Given the high number of SNPs present in the human genome (13 million), it is obvious that many will be identified that could effect the relationship between nutrition and health. Instead of pursuing pathological leads, in particular the effects of these ‘minor’ variants must be studied. Establishing their impact on health, in relation to nutrition, is achieved through cohort-type studies. However, such approaches will fail simply due to a loss of power when multiple minor genetic polymorphisms are involved, all acting on the same physiological process. Nutrition research should not ignore these minor variants, because several subtle changes may together produce large effects, e.g. obesity can both be the result of one pronounced SNP as well as an interplay of many less pronounced variants.

“Nutri-biomics” may contribute in this effort, by describing the mechanisms in terms of kinetic and dynamic models, and verification and validation of these models with genotyping combined with functional analysis on the level of RNA, protein and metabolites.

Conclusions

State-of-the-art technologies are being used to study the effect of dietary habits on health promotion and disease prevention. Not all individuals respond identically to dietary interventions, and meaningful biological interpretation of the generated data is a very complex issue. Intense collaboration between biologists, analytical scientists, statisticians and bioinformaticians is essential in order to gain the most from this new generation of data. To generate interpretable results one must start with a clear and solid question. The amount of data available in public or proprietary databases is increasing rapidly, underlining the central role of bioinformatics to transform raw data into relevant biological knowledge. Furthermore, use of multiple experimental tools and methods will increase the reliability of the results. Advancement of research in computational and analytical science will gradually transform nutrition into a more systematic and hypothesis-driven science. To accelerate and coordinate successful application of high throughput technologies in the field of nutritional research, collaborative actions and networks will be necessary. Already institutions in the U.S. and abroad have developed programs and organizations to address the impact of nutrition on biological systems in a coordinated effort. Only in this manner can the true interaction between an organism and its diet be ascertained.

References

- Ashburner M, Ball CA, Blake JA, Butler H, Cherry JM, Corradi J, Dolinski K, Eppig JT, Harris MA. 2004. Creating the gene ontology resource: design and implementation. *Genome Res* 11: 1425–1433.
- Brazma A. 2001. On the importance of standardisation in life sciences. *Bioinformatics* 17: 113–114.
- Desiere F, German B, Watzke H, Pfeifer A, Saguy S. 2002. Bioinformatics and data knowledge: the new frontiers for nutrition and foods. *Trends Food Sci Technol* 12: 215–229.

- Fields S, Song O. 1989. A novel genetic system to detect protein-protein interactions. *Nature* 340: 245–246.
- Gygi SP, Corthals GL, Zhang Y, Aebersold R. 2000. Evaluation of two-dimensional gel electrophoresis- based proteome analysis technology. *Proc Natl Acad Sci USA* 97: 9390–9395.
- Harvey CB, Hollox EJ, Poulter M, Wang Y, Rossi M, Auricchio S, Iqbal TH, Cooper BT, Barton R, et al: 1998. Lactase haplotype frequencies in Caucasians: association with the lactase persistence/non-persistence polymorphism. *Ann Hum Genet* 62: 215–223.
- Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox, T, Cuff J, Curwen V. 2002. The Ensembl genome database project. *Nucl Acids Res* 30: 38–41.
- Iafate AJ, Feuk L, Rivera, MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C. 2004. Detection of large-scale variation in the human genome. *Nat Genet* 36: 949–951.
- Kothapalli R, Yoder SJ, Mane S, Loughran TP Jr. 2002. Microarray results: how accurate are they? *BMC. Bioinformatics* 3: 22.
- Muller M, Kersten S. 2003. Nutrigenomics: goals and strategies. *Nat Rev Genet.* 4: 315–322.
- Nimgaonkar A, Sanoudou D, Butte AJ, Haslett JN, Kunkel LM, Beggs AH, Kohane IS. 2003. Reproducibility of gene expression across generations of Affymetrix microarrays. *BMC. Bioinformatics* 4: 27.
- Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M. 2004. Large-scale copy number polymorphism in the human genome. *Science* 305: 525–528.