Toxicogenomics (toks´-i-ko-je-nom´-iks): the study of genes and their products important in adaptive response to toxic exposures.

Toxicogenomics: “The Call of the Wild Chip”

Toxicogenomics is a new undertaking in the pursuit of human genes relevant to health risk from environmental toxicants and related stress. Such stress spans a range of dire threats in the human condition from ultraviolet to blue light, from particulate diesel exhaust to complex chemical mixtures in the air, from threats in reused airplane air to threats in herbal remedies, from threats in the food supply from pesticides to threats from genetically modified organisms. This incredible range of interest is embodied in the Environmental Genome Project, sponsored and maintained in the interest of health by the National Institute of Environmental Health Sciences (NIEHS), and about which much has already been written (*1–4*). The aims of the Environmental Genome Project include a characterization of important genetic polymorphisms that alter protein function to the benefit or to the determent of individuals upon exposure to environmental stress. The completion of the human genome sequencing and hopefully its successful annotation is a major spur to accelerated progress in the realm of environmental response genes. The program coordinates a robust suite of extramurally funded research grants, Environmental Health Centers and intramural research projects.

An important part of toxicogenomics is the widespread use of high throughput expressed gene analysis or microarray technology (*5–7*). In an absolutely amazing reversal of our traditional abhorrence of “descriptive” surveys, the scientific community has embraced this latest technology as a relatively cheap, reliable, fast, and effective way to protect populations and specific individuals from environmental risk. Why should we be interested in this approach to gene expression studies that uses high-level image processing, high-fidelity polymerase chain reaction (PCR), and clever hybridization protocols for the purpose of determining gene expression levels by comparison of different RNA populations? In a word—scale. The simultaneous examination of a hundred genes is descriptive; the simultaneous examination by comparison and clustering of tens of thousands of genes is a new way to do science.

Basically oligonucleotides or small pieces of cDNA are spotted onto glass slides (or engineered into computer chips), and then RNAs to be compared are hybridized after labeling with two different fluorochromes: Cy3, which fluoresces red, and Cy5, which fluoresces green. The genes studied may be selected for specific purposes, such as those that are specific to environmental stress, or they may be broad based for gene discovery purposes. Fluorescence appears in the spots (or on the chip) and is detected with image analysis microscopes, usually laser scanning microscopes (or a specialized chip reader). When there is a difference between expression levels in the RNAs being compared, the fluorescent signal with one fluorochrome is greater or less than that of the other. This is detected and quantified. When the intensity with the two fluorochromes is the same (i.e., there is no difference in the two RNAs), then a merged (third) color (yellow) appears. The method detects relative changes in differential expression of populations of RNAs, but it may be combined with other techniques such as real-time PCR (*8*) to obtain quantitative data for selected genes. Depending on the availability of sequence information for the genome of interest and availability of probes from the sequences and the density of individual spots, tens of thousands of expressed genes can be simultaneously examined in a single or a few experiments.

A short parable, though. Suppose one had a large blank wall, and onto this wall one painted tens of thousands of squares, either green, red, or yellow. The squares might be painted in some randomly determined order or by a grand design. Now, suppose your funding agency gave you 100–200 darts labeled “important new research directions” and instructed you to throw them at the wall. Importantly, you have a map of the squares on the wall, which, like the map on the inside cover of a box of assorted chocolates, identifies all of the squares as genes in a particular genome. You throw the darts and then rush to the wall and look for the struck squares that are green or red and declare them “changed gene expression levels.” From those 100 or 200 you may have struck 50 or 60 squares that are “changed gene expression levels.” From the 50 or 60 struck by chance you might become genuinely excited by around 10 or 12 of the “genes” that, because of the course of your own research, you believe to be related to whatever question set you on your path to new knowledge. The rest you would declare interesting, important, and necessary to be cherished for a later time to figure out.

While it is possible that gene expression profiling is described by the foregoing parable, early results indicating specificity of array data suggest that there is true inductive value from this process (*9–11*). It is interesting to note that similar uncertainties existed early in the days of gene sequencing. Importantly, the power of that paradigm was realized when hundreds of researchers working independently of each other on seemingly unrelated genes looked at each other’s sequences and realized that we had much in common with each other and with the species that we study. That was brought about, of course, by widely accessible, accurate searchable databases of sequence information that we all agreed to use (Table 1) and by dramatic increases in the efficiency of DNA sequencing. Can the same approach work for microarray studies? It seems very probable.

Comparison of protocols, tissues, array production methodology, and myriad other variables will lead to an understanding of which technical details alter the results and which are not important. The NIEHS has committed to creating the tools to do this, and it will take the shape of the National Center for Toxicogenomics (NCT) (*24*), which will incorporate the NIEHS Microarray Center. The NCT will be at the NIEHS and will coordinate both extramural and intramural activities with wide access to the toxicology community. An important part of the NCT will be the establishment and maintenance of a toxicology database for microarray results. The NCT will also support the ToxChip, already developed at the NIEHS (*25*). Such a microarray database, GEO (Gene Expression Omnibus), is already up and running at the National Center for Bioinformatics (*13*). There are some problems with the database approach. For example, why would anyone want to enter their data into the program? And how can the quality of the data be assessed before they are released to the public? Here some guidance comes from the other major molecular biology databases. Data can be deposited and held until publication. Peer-reviewed publication could be taken as a surrogate for the legitimacy of the data, and data could be clustered by experimental approach or by the type of question studied. Techniques in tissue-based arrays could be enhanced through international workshops. These might be part of the National Toxicology Program (NTP), which after all has the most complete set of toxicologic response data in the world, with tissue repositories and sophisticated methods for selection of compounds to study.

Results from expressed gene array analyses are very sensitive to the methodology used to establish the initial comparison of RNAs. In some situations it is desirable to determine what genes are expressed in tissues from one physiologic state versus another (*26*,*27*) In these experiments, the way in which the tissue is obtained will determine the value of the data, insofar as other differences will be extant in the tissues besides those of primary interest. In some instances it may be desirable to determine what genes are expressed in the presence or absence of the expression of another gene of interest (*28*,*29*). Such an approach is particularly appealing in cascades or networks of interacting genes such as a signal transduction pathway or a developmental hierarchy (*30*). Here we can choose a cell line, transfect in a plasmid that forces high levels of expression of the gene of interest, and then determine what genes change expression levels. The cell line chosen for such an experiment is important. The gene of interest, if transiently expressed, could lead to some subset of genes that respond. In contrast, if we use stable transfection, we might find gene responses that are secondary to the drug selection system. It could be important that there be a recognizable phenotype which varies between the two states examined, although the genes that determine the phenotype might not be related to the gene of interest. It is important to realize that the fluorochromes vary in their stability, so it is important to label the RNAs both ways; that is, a reciprocal hybridization should be performed in which the Cy3-labeled RNA is labeled with Cy5 and the results compared. Low-abundance RNA may be lost, and subtle differences in RNA in states that are compared may be lost due to poor signal-to-noise ratios. However, low-abundance RNA and small changes in RNA level might be very important. The RNA should not be converted to cDNA probes because this may skew results due to preferential amplification of some RNAs relative to others. Even though highly efficient laser scanners and chip readers have been developed and fairly high-density spotted arrays can be managed, the manner in which the array is produced is very important. Poor spot production or misarranged spot reading can cause artifacts in the results.

Particularly appealing is the opportunity to profile conditions (*9*,*31*). For example, if in a difficult diagnostic differential we could establish that a given set of genes was expressed in one disease state but a different set of genes was expressed in another disease state, then determining which genes are expressed in an unknown case could lead to more informed diagnosis (*9*,*11*). In stratifying patients for treatment, prognosis, and research, such an approach is particularly useful (*32*). Here one needs some way of knowing *a priori* what a given sample is in order to determine the gene expression profile; but if that can be done with confidence, what is the value in the end of knowing the profile? In this circumstance we might hope to use a prospective approach that allows elucidation of the correct diagnosis after the fact by the behavior of the disease and with appropriately selected individuals whose RNA can be examined. Moreover, once the analytical method is established, we might expect that it would be rapid and highly reproducible.

A logical extension of the approach of expressed gene profiling is to study physiologic response to environmental stress, in particular toxic exposure, and to determine for any given class of toxicants what genes alter expression in stereotypical ways in a standardized exposure protocol, either of intact test animals or particular cell lines. In this way it seems entirely possible to create a toxicant class-specific profile or fingerprint of expressed genes (*25*,*34*–*36*). The appeal is that then we need only to repeat the exposure protocol for an unknown chemical and declare it dangerous or safe depending on the genes whose expression is altered in the exposure protocol. Obviously the quality of the resulting profiles and their usefulness will depend on the exposure protocol and its appropriateness to the supposed class of the toxicant. Moreover, it will depend on the cluster of toxicants included in a class. If the class of toxicants is too broad, too many genes would be expressed and the genes would not be specific; if the class of toxicants is too narrow, there could be almost as many classes as chemicals to be studied and the application of the profile to the chemical would become hopelessly ad hoc. Indeed microarrays that allow the examination of toxic responses are already available and under close scrutiny (*25*).

There are issues that current approaches cannot fully address. In complex signal transduction pathways, the activation of transcription factors is frequently the result of protein–protein interactions and not solely the result of genetic regulation. For example, in the

Table 1. Web-based resources for toxicogenomics.

Tool	Related operations	National and international databases	Reference
Overview			(*12*)
Gene expression profiling	cDNA microarrays	Gene Expression Omnibus	(*13*)
Transcriptome	Expressed sequence tags	UniGene	(*14*)
Genomics	Genebank	LocusLink	(*15*)
Gene mapping	Anonymous markers	GeneMap (human)	(*16*)
	Radiation hybrid maps		
	Polymorphisms		
	Species-specific	RatMap (rat)	(*17*)
		Rat Genome Database (rat)	(*18*)
		FlyBase (*Drosophila*)	(*19*)
		Mouse Genome Informatics (mouse)	(*20*)
Proteomics	Molecular modeling	Protein Data Bank	(*21*)
	Crystallographic coordinates	SwissProt	(*22*)
		Molecular Modeling Database	(*23*)
Phenotyping	Tissue arrays	Not organized	

Sonic hedgehog–Patched–GLI pathway (*36,37*) the transcription factors GLI1 and GLI3 are tethered to cytoskeletal elements. In the presence of signaling they are released to be transported to the nucleus as a result of an activation mechanism that is not well worked out, but certainly involves the cleavage of GLI3 and probably phosphorylaton of the binding proteins. Modification of one of these binding proteins, Slimb, may be critical to the activation of the transcription factors. Precisely how this occurs and which other helper proteins are involved in the process will require the same type of high throughput large-scale technology currently being applied to expressed gene sets. Technologies are currently being developed that will allow this (*38*), and they no doubt will be important in our study of environmental stress response and attempts to develop toxic signatures for unknown compounds. Even at the genetic level, not all regulation will occur at the level of RNA transcription and may not be easily detected by subtraction type strategies. For example, the regulation of protein production through translation control is an important aspect of several genetic pathways. In the case of *Tra 2* in *Caenorhabditis elegans* and *GLI1* in humans, the 3´UTR in the transcribed RNA contains an element that binds a regulatory protein(s) which leads to shortening of the poly A(+) tail and results in lower levels of protein production (*39–41*). This mechanism and the regulatory binding proteins are conserved from the worm to vertebrates. Clearly, the binding proteins and the elements they bind to are targets of environmental stress. Microarray data cannot help to elucidate regulation that results from subcellular compartmentalization of molecules, as in the case of p53 or cyclins. Where regulation results from reversible phosphorylation or from ubiquination, the gene expression data will not be helpful.

Arguably the largest, most important environmental challenge facing the American population in the next 25 years is air pollution. Purely as a quantitative matter, the volume of environmental stressors that we expose ourselves to, the sheer surface area of the exposure in our bodies (i.e., the total airway surface area), and the complexity of the chemistry behind it should be compelling enough. When we consider the extraordinary growth of the sources of air pollution in this country, it is hard to imagine a more important environmental health issue. A wide variety of serious pulmonary diseases have not been well characterized, and there is a major need to establish paradigms in health studies that cross over the traditional emphasis on chemistry of exposure versus the study of pathology of lung diseases. This is particularly true of childhood morbidity. In many children's hospitals pulmunary diseases such as asthma and cystic fibrosis represent leading causes of admission. It is important to consider that events in childhood can be precursors of adult chronic disease, greatly increasing the burden on society. Because these events have environmental components, we should establish ways to study pulmonary disease in the context of the complex chemistry of air pollution. This will require sophisticated epidemiology, interventional epidemiology, and robust studies in pulmonary biology. The new genomic information and toolboxes may lead to further progress.

It may be possible to screen biological samples obtained from workers at Superfund sites for the adverse effects of exposure to compounds present in the site. Although the Superfund project has made important advances in public health in the United States, we have a great deal to learn about the effects on workers at these sites. Moreover, we are now beginning to realize the full magnitude of these sites. As we begin to deal with seriously contaminated sites in the Department of Defense and the Department of Energy, the magnitude of true cleanup problems will strike home. We should be prepared to understand the effects on workers performing these cleanups. Once we know a suite of genes that are important in health outcomes at such sites, it may be possible to use real time PCR (*8*) on the spot as a signature of both exposure and risk.

We must continue to grapple with the question of how the mass of data that will be generated in the next five years can be translated into experiments which will help regulators. One need of toxicogenomics involves the development of sound policy that must be based in good science. The challenge is to first create the good science, then to generate the discussion that informs policy. The standards of analysis, toxicogenomics databases, and public discussion can provide a platform that supports the creation of good science. The challenge will be to figure out how.

Many people who are worried about environmental stress wonder if a given exposure in a given setting will pose a threat to an individual's health. What damage occurs if a cloud of diesel exhaust blows in your face? What is the recovery phase from the exposure? What is the health consequence? How frequently does the exposure need to occur before these health effects are permanent? The answers will depend, in part, on the effects of polymorphisms in the response genes, and these are not yet fully known. It will be important to coordinate the efforts of the Environmental Genome Program to find the meaningful polymorphisms with gene expression profiling. Exposure effects present a longitudinal problem: If the relevant tissue could be obtained from the relevant population, the throughput potential of the new toxicogenomics technologies could allow the dimension of time in accumulated exposure risk to health to be addressed. The challenge is to determine how.

There are large human cohort studies under way nationally in several important areas. For example, the National Cancer Institute and the NIEHS are accumulating long-term data from tens of thousands of agricultural workers with respect to pesticide exposure. Can the tissue and data from the pesticide study be linked to gene expression studies in a meaningful way? The throughput potential of the technologies discussed above should allow this, but we must figure out how to do it.

There is a great deal of evidence that points to the effects of environmental stress on infectious disease outcomes. It seems axiomatic that there are important interactions of toxic exposure with pathogens that will effect health outcomes. Can expressed gene array data, gene polymorphism data, or human sequencing lead to identification of overlapping gene pathways in pathogen response?

The completion of the *Drosophila* genome led to a number of exciting intellectual developments. Among these was a change in biomedical research paradigms toward consortia or engineering models of cooperative effort. This approach has been used in the past, but never, I believe, on such a scale. Another exciting development was the manner in which the genome was annotated and the information that flowed from the annotation (*42*). This approach to a massive data set may inform the growing field of toxicogenomics. Can there be an environmental response gene annotation emphasis on the human genome project data?

As we look back over the past 12 or 15 years of genetics, we ask ourselves what we have learned about large efforts and the payoff they can have for important advances and for the growth of the general body of knowledge. What tools have survived the test of time (even though relatively short) and the test of utility? The pioneering efforts of sequence databases, protein structure data, and the accessible tools that allow meaningful use of the data are at the forefront. Rapid emergence of computing tools that do not require intimate knowledge of computer architecture to operate them have allowed important discovery, new knowledge, and, importantly, new experimental designs to flourish. The expressed sequence tag (EST) programs have not only facilitated physical mapping of chromosomes but have allowed innovative use of genome-wide information. The involvement of the National Center for Bioinformatics in the creation of a database of unique genes in the EST pool [UniGene (*14*)] has provided highly accessible information and clone sources to investigators worldwide. It should be anticipated that computing resources important to the range of molecular techniques in toxicogenomics (sequencing, polymorphisms, phenotyping, expressed gene microarrays, proteomics) will become available and widely used. The development of tissue arrays should offer a new chance for the NTP to provide a leap forward in our understanding of non-cancer end points of environmental stress. The NTP has a huge collection of tissue and data from animals exposed to nominated compounds that could be used to determine the longitudinal response of environmentally sensitive genes to exposure. The Environmental Genome Project will also develop such a repository. The Agricultural Health Study will likely develop a repository that could also be used in this way. I hope that the information and tissues can be made broadly available to the scientific community so that they can be fully used to the benefit of public health.

As the United States continues to struggle with a toxic legacy and with an ever-increasing burden of environmental stress, especially from air pollution, new methodologies to assess risk to both populations and individuals must be put in place for our protection. How will the scientific community respond to this challenge? In what ways will the genetic revolution of the past several decades contribute to this vital cause? We may hope that as toxicogenomics approaches maturity, the concentration of a suite of tools and the management of large data sets can lead to a point where meaningful science will guide public health policy and minimize personal health risk.

Philip M. Iannaccone
Department of Pediatrics
Northwestern University Medical School
The Children's Memorial Institute for Education and Research
Chicago, Illinois
E-mail: pmi@nwu.edu

References and Notes

1. Sharp RR, Barrett JC. The environmental genome project: ethical, legal, and social implications. Environ Health Perspect 108:279–281 (2000).
2. Guengerich FP. The Environmental Genome Project: functional analysis of polymorphisms. Environ Health Perspect 106:365–368 (1998).
3. Shalat SL, Hong JY, Gallo M. The Environmental Genome Project. Epidemiology 9:211–212 (1998).
4. Environmental Genome Project. Available: http://www.niehs.nih.gov/envgenom/ [last update 4 October 2000].
5. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, et al. Functional discovery via a compendium of expression profiles. Cell 102:109–126 (2000).
6. Afshari CA, Nuwaysir EF, Barrett JC. Application of complementary DNA microarray technology to carcinogen identification, toxicology, and drug safety evaluation. Cancer Res 59:4759–4760 (1999).
7. Schena M, Heller RA, Theriault TP, Konrad K, Lachenmeier E, Davis RW. Microarrays: biotechnology's discovery platform for functional genomics. Trends Biotechnol 16:301–306 (1998).
8. Bustin SA. Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays. J Mol Endocrinol 25:169–193 (2000).
9. Kononen J, Bubendorf L, Kallioniemi A, Barlund M, Schraml P, Leighton S, Torhorst J, Mihatsch MJ, Sauter G, Kallioniemi OP. Tissue microarrays for high-throughput molecular profiling of tumor specimens. Nat Med 4:844–847 (1998).
10. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, et al. Molecular portraits of human breast tumours. Nature 406:747–752 (2000).
11. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403:503–511 (2000).
12. NCBI Databases. Available: http://www.ncbi.nlm.nih.gov/Database/index.html [cited 7 December 2000].
13. Gene Expression Omnibus. Available: http://www.ncbi.nlm.nih.gov/geo/ [cited 7 December 2000].
14. UniGene. Available: http://www.ncbi.nlm.nih.gov/UniGene/ [cited 7 December 2000].
15. LocusLinks. Available: http://www.ncbi.nlm.nih.gov:80/LocusLink/ [cited 7 December 2000].
16. A New Gene Map of the Human Genome. The International RH Mapping Consortium. Available: http://www.ncbi.nlm.nih.gov/genemap/ [cited 7 December 2000].
17. RatMap (rat). Available: http://ratmap.gen.gu.se/ [cited 7 December 2000].
18. Rat Genome Database (rat). Available: http://rgd.mcw.edu/ [cited 7 December 2000].
19. FlyBase, A Database of the Drosophila Genome. Available: http://flybase.bio.indiana.edu/ [cited 7 December 2000].
20. Mouse Genome Informatics. Available: http://www.informatics.jax.org/ [cited 7 December 2000].
21. Protein Data Bank. Research Collaboratory for Structural Bioinformatics. Available: http://www.rcsb.org/pdb/ [cited 7 December 2000].
22. EMBL Outstation European Bioinformatics Institute SwissProt. Available: http://www.ebi.ac.uk/swissprot/ [cited 7 December 2000].
23. Molecular Modeling Database. Available: http://www.ncbi.nlm.nih.gov:80/Structure/MMDB/mmdbpub.shtml [cited 7 December 2000].
24. Press Release: New Center to Look for Precise Step When a Cell Tips Toward Death, Disease. Available: http://www.niehs.nih.gov/oc/news/toxgen.htm [cited 7 December 2000].
25. Medlin JF. Innovations: Timely toxicology. Environ Health Perspect 107:A256–258 (1999).
26. Mariadason JM, Corner GA, Augenlicht LH. Genetic reprogramming in pathways of colonic cell maturation induced by short chain fatty acids: comparison with trichostatin A, sulindac, and curcumin and implications for chemoprevention of colon cancer. Cancer Res 60:4561–4572 (2000).
27. Zimmermann J, Erdmann D, Lalande I, Grossenbacher R, Noorani M, Furst P. Proteasome inhibitor induced gene expression profiles reveal overexpression of transcriptional regulators ATF3, GADD153 and MAD1. Oncogene 19:2913–2920 (2000).
28. Coller HA, Grandori C, Tamayo P, Colbert T, Lander ES, Eisenman RN, Golub TR. Expression analysis with oligonucleotide microarrays reveals that MYC regulates genes involved in growth, cell cycle, signaling, and adhesion. Proc Natl Acad Sci USA 97:3260–3265 (2000).
29. Feng X, Jiang Y, Meltzer P, Yen PM. Thyroid hormone regulation of hepatic genes in vivo detected by complementary DNA microarray. Mol Endocrinol 14:947–955 (2000).
30. White KP, Rifkin SA, Hurban P, Hogness DS. Microarray analysis of Drosophila development during metamorphosis. Science 286:2179–2184 (1999).
31. Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, Van de Rijn M, Waltham M, et al. Systematic variation in gene expression patterns in human cancer cell lines. Nat Genet 24:227–235 (2000).
32. Moch H, Schraml P, Bubendorf L, Mirlacher M, Kononen J, Gasser T, Mihatsch MJ, Kallioniemi OP, Sauter G. High-throughput tissue microarray analysis to evaluate genes uncovered by cDNA microarray screening in renal cell carcinoma. Am J Pathol 154:981–986 (1999).
33. Bartosiewicz M, Trounstine M, Barker D, Johnston R, Buckpitt A. Development of a toxicological gene array and quantitative assessment of this technology. Arch Biochem Biophys 376:66–73 (2000).
34. Fornace AJ Jr, Amundson SA, Bittner M, Myers TG, Meltzer P, Weinsten JN, Trent J. The complexity of radiation stress responses: analysis by informatics and functional genomics approaches. Gene Expr 7:387–400 (1999).
35. Nuwaysir EF, Bittner M, Trent J, Barrett JC, Afshari CA. Microarrays and toxicology: the advent of toxicogenomics. Mol Carcinog 24:153–159 (1999).
36. Villavicencio EH, Walterhouse DO, Iannaccone PM. The sonic hedgehog-patched-Gli pathway in human development and disease. Am J Human Genet 67:1047–1054 (2000).
37. Walterhouse DO, Yoon JW, Iannaccone PM. Developmental pathways: Sonic hedgehog-Patched-GLI. Environ Health Perspect 107:167–171 (1999).
38. Ge H. UPA, a universal protein array system for quantitative detection of protein-protein, protein-DNA, protein-RNA and protein-ligand interactions. Nucleic Acids Res 28:e3 (2000).
39. Graves LE, Segal S, Goodwin EB. TRA-1 regulates the cellular distribution of the tra-2 mRNA in *C. elegans*. Nature 399:802–805 (1999).
40. Saccomanno L, Loushin C, Jan E, Punkay E, Artzt K, Goodwin EB. The STAR protein QKI-6 is a translational repressor. Proc Natl Acad Sci USA 96:12605–12610 (1999).
41. Jan E, Yoon JW, Walterhouse D, Iannaccone P, Goodwin EB. Conservation of the *C.elegans* tra-2 3'UTR translational control. Embo J 16:6301–6313 (1997).
42. Rubin GM, Yandell MD, Wortman JR, Gabor Miklos GL, Nelson CR, Hariharan IK, Fortini ME, Li PW, Apweiler R, Fleischmann W, et al. Comparative genomics of the eukaryotes. Science 287:2204–2215 (2000).