Minireview

Host-pathogen studies in the post-genomic era

Paul Kellam

Address: Wohl Virion Centre, Department of Molecular Pathology, Windeyer Institute, University College London, London, W1P 6DB, UK.
E-mail: p.kellam@ucl.ac.uk

Published: 4 August 2000

Genome ***Biology*** 2000, **1(2)**:reviews1009.1–1009.4

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2000/1/2/reviews/1009

Abstract

Several studies are starting to show the power of DNA microarrays to identify interactions between animal hosts and their pathogens, and have revealed interesting correlations between host responses to different infectious agents.

Introduction

Post-genomic research is now firmly established as a major scientific discipline in the new millennium. The first working draft of the human genome is now available, and predictions of the human gene content will be available soon. Virology has been in the post-genomic era since 1977, with the sequencing of the øX174 genome [1], and GenBank now holds more than 1,000 complete viral genomes. Bacteriology has also been post-genomic since completion of the *Haemophilus influenzae* genome sequence in 1995 [2]. Parallel to the sequencing of large genomes has been the rapid development of methods for studying the expression of the information they encode. With the advent of DNA microarray and chip technologies, gene expression can now truly be explored on a 'genome scale' [3]. In research into infectious disease, we are now rapidly approaching the time when it will be possible to study gene expression of both host and pathogen at the whole-genome level. Realizing the promise of the post-genomic era is, however, largely dependent on harnessing expertise from all aspects of biology, underpinned in an integrative manner by computational biology. This is particularly relevant in host-pathogen studies, which, as well as 'post-genomic' scientists, require virologists, bacteriologists, parasitologists, immunologists and cell biologists.

Ways of studying gene expression

Large-scale expression studies now mean it is possible to define an organism's phenotypic state in any given condition according to which genes are expressed. This has been defined as the 'transcriptome'. Large-scale gene-expression mapping using arrays is motivated by the premise, based on the central dogma of molecular biology, that the functional state of the organism is largely determined by the information carried by its expressed genes. In reality, things are not that simple, as the relationship between the absolute amounts of some proteins and the level of their corresponding transcripts is more complex than a simple linear one. Nevertheless, much can be gained from this type of study.

There are several different methods of measuring gene expression, including quantitative RT-PCR, serial analysis of gene expression (SAGE), Affymetrix-type oligonucleotide microarray 'chips' and DNA-based microarrays (Table 1). I concentrate here on the use of microarrays (see Box 1).

One current problem with the different methods of quantifying gene expression is the lack of systematic assessment of the comparability of results. Each method tends to produce different representations of a gene expression level. It is widely acknowledged that experiments using the same samples but a range of methods are urgently required in order to understand the relative merits of each system [4]. This is important, as it is unlikely that one method of measuring gene expression will be universally accepted. Over time, however, there may be a gradual shift to the use of one broad type of methodology, as occurred with the widespread preference for Sanger dideoxy-chain terminator sequencing over Maxam and Gilbert chemical

Table 1

DNA array terminology

Type of ORF probe	Solid support	Name
Synthesized oligonucleotide	Glass	Affymetrix
PCR product or cloned DNA	Glass	Microarray
	Nylon	Macroarray

When 'DNA array' is referred to in the text, it encompasses all types of array included in the table. ORF, open reading frame.

degradation sequencing. Currently, DNA arrays seem to be the method of choice for monitoring of large-scale gene expression.

Host and pathogen gene expression

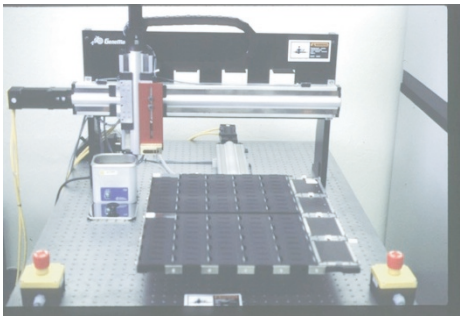
Despite being still in their infancy, DNA arrays have been used to study host and pathogen gene expression profiles for four viruses - human cytomegalovirus (HCMV) [5,6], human herpesvirus 8 (HHV8) (R.G. Jenner, M. Mar Albà, C. Boshoff, and P. Kellam, unpublished observations), human immunodeficiency virus type-1 (HIV-1) [7] and human papillomavirus type 31 (HPV31) [8], as well as two bacterial pathogens - *Listeria monocytogenes* [9] and *Salmonella* [10]. Two studies focused on the complete gene expression profiles of the pathogen ([5] and our unpublished observations) with the rest focusing on the expression of subsets of host genes (Table 2). Most of these studies experienced the problems inherent in dealing with the masses of data produced with DNA arrays and confined their analysis to listing genes that were up- or downregulated. Our study of HHV8 gene expression used cluster analysis [11], a tool for rationalizing gene expression patterns into groups of coordinately expressed genes. Cluster analysis has been used to group genes involved in similar processes and provides an insight into the biology of the system studied [12,13]. In our study of HHV8, this analysis provided further information on the coordination of viral gene expression during replication.

Common patterns of host gene expression in response to different pathogens are difficult to determine from the current studies. This is mainly due to the different systems used and inconsistencies in the annotation of host genes. Many responses of the host to different pathogens are already known [14], but a more comprehensive whole-genome analysis may have far-reaching effects on understanding the pathogenesis of different infections. From the five studies focusing on the host response (Table 2), it is possible to determine a small number of genes that are consistently detected as up- or down-regulated (Table 3). Infection with both bacterial pathogens upregulates expression of the chemokines interleukin-8 (IL-8), GROβ (macrophage inflammatory protein 2α, MIP2α) and leukemia inhibitory factor (LIF). IL-8 is released by several cell types in response to an inflammatory stimulus and is a chemoattractant for neutrophils, basophils and T cells. GROβ is also known to be expressed at sites of inflammation, and LIF is able to induce hematopoietic differentiation of myeloid progenitor cells. Therefore, expression of these chemokines is consistent with the need to attract and activate leukocytes to bacterially infected tissues.

Box 1

A microarray is a surface that contains representations of each open reading frame (ORF) of a sequenced and annotated genome. Of the several available formats, the most commonly used in academic labs (developed by Patrick Brown and colleagues at Stanford University) consists of a microscope slide whose surface displays a matrix of printed spots, each spot containing a PCR-derived amplicon that corresponds to all or part of an ORF of the sequenced genome. Thus, each ORF of the genome is represented on the array as a separate spot, its location designated by its matrix coordinates.

One principal innovation in gene-expression profiling involved the introduction of two-color hybridization. This method employs two populations of cDNAs that have been differentially labeled with two different fluorochromes – the cDNAs usually having been derived from RNA prepared from the same organism cultivated under or exposed to, two contrasting conditions. Equal masses of the two differentially labeled populations of cDNAs are combined, applied to the array surface and allowed to hybridize to the corresponding ORF-specific targets. The array is then scanned and the intensity of each label for each ORF-specific spot is quantitated. These values are compared, yielding ratios that serve as a measure of the relative expression levels of each ORF for the two tested conditions.



Other microarray systems and methods, such as those developed by Affymetrix, really differ only in detail (for example, the oligonucleotides representing each ORF are synthesized *in situ* on the solid support, rather than being spotted onto glass slides or nylon membranes), but the underlying principles of experimental design remain the same.

Table 2

Host and pathogen DNA array studies

Cell type	Pathogen	Array	Genes	Array type and support	Experiment	Reference
Human foreskin fibroblasts	Human cytomegalovirus	Pathogen	207 genes	Oligonucleotides on glass slides	Virus gene expression during lytic replication	[5]
Human B-cell*	Human herpesvirus 8	Pathogen	120 genes	PCR products on nylon	Virus gene expression during latent and lytic replication	‖
Human T-cell†	Human immuno-deficiency virus type 1	Host	1,500 genes	PCR products on glass slides	Host gene expression during 72 hours of virus infection	[7]
Normal human keratinocytes	Human papillomavirus type 31	Host	7,075 genes	PCR products on glass slides	Host gene expression following transfection of the viral genome	[8]
Human foreskin fibroblasts	Human cytomegalovirus	Host	6,600 genes	Affymetrix microarrays	Host gene expression during 24 hours of virus infection	[6]
Human colorectal‡ and colon§ epithelial cells	Salmonella dublin	Host	4,300 genes	PCR products on nylon	Host gene expression during 20 hours of bacterial infection	[10]
Human monocytes¶	*Listeria monocytogenes*	Host	6,800 genes 18,367 genes 588 genes	Affymetrix microarrays PCR products on nylon	Host gene expression during 2 hours of bacterial infection	[9]

*Primary effusion lymphoma cell line BC-3; †T-cell lymphoma cell line CEMCCRF; ‡HT-29 cells; §T84 cells; ¶THP-1 cells. ‖R.G. Jenner, *et al*., unpublished observations.

Tyrosine phosphorylation and interaction of signaling proteins are the foundation of many signaling pathways. General control of tyrosine phosphorylation of signaling molecules is accomplished through the action of phosphotyrosine phosphatases (PTPs). It is necessary for cells that both protein PTPs and protein tyrosine kinases maintain their physiological balance in order to sustain normal regulation of events dependent on phosphorylated tyrosine residues. Inhibitors of certain PTPs have been shown to inhibit the growth of the protozoan pathogen *Leishmania* [15], owing, in part, to increased sensitivity of host cells to interferon-γ stimulation. On the other hand, inhibitors of PTP have also been shown to activate the replication of HIV-1 by both NFκB-dependent and -independent pathways [16]. Taken together, this suggests a reason for pathogen modulation of different PTP genes as indicated in Table 3 and indicates that pathogens may exploit PTPs during their replicative cycle.

It will be interesting to determine whether the host produces a consistent broad response to viruses or bacterial infections, or if the host is able to discriminate and tailor its response to different types of virus - for example, poliovirus, with a single-stranded mRNA sense genome, compared with herpesviruses, with double-stranded DNA genomes - and bacteria - for example, Gram-positive versus Gram-negative. In addition, post-genomic research may help to answer complex questions

Table 3

Common genes up- or down-regulated during infection by bacteria and viruses

Gene	GenBank accession number	*Listeria*	*Salmonella*	CMV	HIV-1	HPV31
Interleukin-8	M28130	U	U	-	-	-
GROβ/macrophage inflammatory protein 2α	M57731	U	U	-	-	-
Leukemia inhibitory factor	X13967	U	U	-	-	-
Receptor phosphotyrosine phosphatase, PCP-2	X97198	-	-	-	-	D
Type IVA phosphotyrosine phosphatase	AA504327	-	-	-	D	-
Phosphotyrosine phosphatase-BAS Type 1	D21209	-	-	D	-	-
Phosphotyrosine phosphatase/MKP-1	X68277	U	-	-	-	-
Phosphotyrosine phosphatase/PAC-1	L11329	U	-	-	-	-
Interferon α-inducible p27 protein	X67325	-	-	-	U	D

U; upregulated, D; downregulated; -, not present in array data. The bacteria and viruses are as in Table 2.

about pathogen persistence. For example, the quite closely related yellow fever virus and hepatitis C virus result in very different pathologies, yellow fever virus producing an acute, sometimes fatal, infection, whereas hepatitis C virus forms a long-term persistent infection that ultimately leads to liver cancer. Also, no attempts have yet been made to incorporate host and pathogen genes into the same DNA array to determine the coordinated interactions between host and pathogen. These sorts of studies are likely to reveal much new information and may ultimately lead to better targeted anti-infective therapeutics and enhanced vaccination strategies.

Data analysis and integration

To address many questions about host-pathogen interactions, methods of data analysis and integration must improve. Post-genomic studies, by their very nature, produce vast amounts of data. The true potential of methods such as DNA arrays will, however, only be realized by careful data management and bioinformatics analysis. A new breed of biologist is emerging who not only understands his or her particular biological system but is also computer literate and able to handle, analyze and conceptualize vast amounts of biological data. This has led to the realization that carefully designed and maintained databases are now a must for many laboratories, and data-warehousing of additional related information is likely to be essential for discovering underlying patterns and relationships in the data.

Most DNA array laboratories have in-house databases for their own array experiments. Of greater value would be public expression databases such as ArrayExpress, envisaged by the European Bioinformatics Institute [17,18], and the National Cancer Institute's ArrayDB [19]. These will function as repositories for array data analogous to the sequence databases EMBL, GenBank and DDJB. In the future, it is likely that publication of expression data in journals will require the submission of data to a public expression database and the assignment of an accession number prior to publication, again analogous to submission of new sequence data. Gene expression data are at present far from suitable for such databases, however. In comparison to DNA sequence or protein structure data, gene expression data are stored mainly as unstructured flat-files with no uniform standards of data reporting [4]. Different methodologies report different types of quantitation of gene expression, and the relationships between the different methods are not yet fully understood. This has led the array community to propose a minimum information standard and data format for expression data to facilitate the construction of a public database [4,18,19].

Such databases will be essential to enable detailed cross-comparison between different cellular expression patterns under various conditions. As outlined above, this is important for host-pathogen studies, in which integrated analyses of normal and infected cells, pathogen-expressed genes and host immune system genes will need to be compared. Integration of other post-genomic information, such as proteomics data, will also be needed. Furthermore, the eventual integration of gene-specific information from other databases in regard to structure, function, and biological process, and of specialist data relating to the pathogens, will equip biologists with the information and, hopefully, sufficient understanding of host-pathogen interactions, to generate further testable hypotheses.

References

1. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocombe PM, Smith M: **Nucleotide sequence of bacteriophage phi X174 DNA.** *Nature* 1977, **265:**687-695.
2. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM *et al.*: **Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd.** *Science* 1995, **269:**496-512.
3. **The chipping forecast**. *Nat Genet* 1999, **21(Suppl)** [http://www.nature.com/ng/chips_interstitial.html].
4. Aach J, Rindone W, Church GM: **Systematic management and analysis of yeast gene expression data.** *Genome Res* 2000, **10:**431-445.
5. Cambers J, Angulo A, Amaratunga D, Guo H, Jiang Y, Wan JS, Bittner A, Frueh K, Jackson MR, Peterson PA, Erlander MG, Ghazal P: **DNA microarrays of the complex human cytomegalovirus genome: profiling kinetic class with drug sensitivity of viral gene expression.** *J Virol* 1999, **73:**5757-5766.
6. Zhu H, Cong J-P, Mamtora G, Gingeras T, Shenk T: **Cellular gene expression altered by human cytomegalovirus: global monitoring with oligonucleotide arrays.** *Proc Natl Acad Sci USA*. 1998, **95:**14470-14475.
7. Giess GK, Bumgarner RE, An MC, Agy MB, Van't Wont AB, Hammersmark E Carter VS, Upchurch D, Mullins JI, Katze MG: **Large-scale monitoring of host cell gene expression during HIV-1 infection using cDNA microarrays.** *Virology* 2000, **266:**8-16.
8. Chang YE, Laimins LA: **Microarray analysis identifies interferon-inducible genes and Stat-1 as major transcriptional targets of human papillomavirus type 31.** *J Virol* 2000, **74:**4174-4182.
9. Cohen P, Bouaboula M, Bellis M, Baron V, Jbilo O, Poinot-Chazel C, Galiegue S, Hadibi E-H, Casellas P: **Monitoring cellular responses to *Listeria monocytogenes* with oligonucleotide arrays.** *J Biol Chem* 2000, **275:**11181-11190.
10. Eckmann L, Smith JR, Housley MP, Dwinell MB, Kagnoff MF: **Analysis of high density cDNA arrays of altered gene expression in human intestinal epithelial cells in response to infection with the invasive enteric bacteria *Salmonella*.** *J Biol Chem* 2000, **275:**14084-14094.
11. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns.** *Proc Natl Acad Sci USA* 1998, **95:**14863-14868.
12. 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* 2000, **403:**503-511.
13. Chu S, DeRisi J, Eisen M, Mulholland J, Botstein D, Brown PO, Herskowitz I: **The transcriptional program of sporulation in budding yeast.** *Science* 1998, **282:**699-705.
14. Luster AD: **Chemokines - chemotatic cytokines that mediate inflammation.** *N Engl J Med* 1998, **338:**436-445.
15. Olivier M, Romero-Gallo B-J, Matte C, Blanchette J, Posner BI, Trembley MJ, Faure R: **Modulation of interferon-γ induced macrophage activation by phophotyrosine phosphatases inhibition.** *J Biol Chem* 1998, **273:**13944-13949.
16. Barbeau B, Bernier R, Dumais N, Braind G, Olivier M, Faure R, *et al.*: **Activation of HIV-1 long terminal repeat transcription and virus replication via NF-κB dependent and independent pathways by potent phosphotyrosine phosphatase inhibitors, the peroxovanadium compounds.** *J Biol Chem* 1997, **272:**12968-12977.
17. Abbott A: **Bioinformatics institute plans public database for gene expression data.** *Nature* 1999, **398:**646.
18. **The ArrayExpress database** [http://www.ebi.ac.uk/arrayexpress/].
19. Ermolaeva O, Rastogi M, Pruitt KD, Schuler GD, Bittner ML, Chen Y, Simon R, Meltzer P, Trent JM, Boguski MS: **Data management and analysis for gene expression arrays.** *Nat. Genet* 1998, **20:**19-23.