Anyone who has taken an undergraduate virology course is familiar with subject matter focused on the structure of viral genomes and the molecular events associated with multistep viral life cycles. The field of virology has done a remarkable job of characterizing and categorizing viruses and defining the steps of viral attachment, entry, replication and release. Moreover, an understanding of viral protein function has paved the way for the development of antiviral drugs that target viral enzymatic activities. However, many of these drugs function poorly at best, and the virus-centric approach has not proved to be well suited for deciphering the complex and multifaceted virus–host interactions that underlie viral recognition, innate immune signalling and disease outcome. In the past decade, tools have become available to chart a new course, one directed at obtaining comprehensive systems-level views of the host response and the interplay between virus and host.

Systems virology is a phrase coined to describe the application of systems biology approaches to the field of virology. Systems biology is highly interdisciplinary in character, requiring the combined talents of biologists, mathematicians and computer scientists, and the goal of the field is to gain comprehensive understanding of biological systems. In the case of systems virology, these biological systems can range from virus-infected cells to tissues to whole organisms. Systems-level analyses use high-throughput technologies to measure system-wide changes in biological components such as DNA, RNA, proteins and metabolites, and are dependent on the quality of both the resulting data sets (which are often noisy) and the subsequent data integration and modelling. Ideally, high-throughput data derived from these and other measurements are integrated and analysed using mathematical algorithms to generate predictive models of the system. When a model has been developed, subsequent experimental perturbations of the system (for example, the use of viral mutants or targeted inhibition of host genes or pathways) are used to yield refinements to the model and to increase its predictive capacity.

This holistic, host-directed approach stands in contrast to the more traditional reductionist approaches that focus on a pre-determined small set of molecules (genes, proteins or metabolites). Although systems-level (or discovery-based) analyses are often criticized for not being hypothesis-driven, these analyses are increasingly being acknowledged as potent hypothesis generators. Moreover, for dynamic systems such as those involved in the host response to viral infection, systems-level analyses are considered the only way to understand emergent properties — that is, properties or biological outcomes that cannot be predicted by an understanding of the individual parts of a system alone, but rather become apparent only with knowledge of the specific organization of, and interactions between, components. Because of this, systems virology is an essential and synergistic complement to traditional virology approaches.

This Review focuses on the host response to virus infection and discusses the evolution and significant advances in systems virology.
findings of systems virology, including the identification of gene expression signatures that are predictive of viral pathogenesis and vaccine efficacy, insights into how viruses disrupt cellular metabolism, and the mapping of virus–host interactomes. These accomplishments did not come from a single experiment or study, but rather from a body of work undertaken over several years by different investigators. The field has seen a progression from genomics-based approaches to measurements of proteins and metabolites, and has also embraced the analysis of host genetic variation as a means to better understand disease processes rather than viewing such variation as a source of frustration. Moving forwards, systems virology must also embrace computational approaches that are capable of integrating this information to construct robust virus–host interaction models that incorporate multiple dimensions and scales. We cite examples of studies that are moving in this direction and outline what the next phase of systems virology must encompass to reach its full potential.

**Gene expression signatures**

With the completion of the human genome project and the advent of microarrays capable of measuring RNA transcripts at a genome-wide scale, the first systems-level analyses became a reality. Over the past 12 years, DNA microarrays have evolved from hundreds of cDNAs spotted on nylon membranes to glass slides containing high-density oligonucleotides that encompass entire genomes (BOX 1). Microarrays (then covering only 1,500 human genes) were first used in virology to evaluate the changes in cellular gene expression that occurred in a CD4 T cell line infected with HIV. Since then, the use of microarrays to evaluate changes in host gene expression in response to viral infection has become commonplace. In many cases, these studies remain small and narrowly focused, and although they provide glimpses into global responses, low sample numbers make it difficult to determine their reproducibility. In addition, these studies do not provide the robust data sets that are needed for computational modelling, which could result in deep insights into system architecture or behaviour. This shortcoming has perhaps fed scepticism regarding the ability of genome-wide expression profiling to yield transformative discoveries. Below, we provide examples of more comprehensive studies resulting in genomic signatures that have increased our understanding of both viral pathogenesis and the characteristics of the host response required for immune protection.

**Signatures of highly pathogenic respiratory viruses.** Influenza virus is well known for its ability to rapidly evolve new variants through genetic mutation and genome reassortment, yielding strains that can vary widely in virulence and transmissibility. Most strains cause mild respiratory disease, whereas others, such as the 1918 pandemic virus and highly pathogenic avian H5N1 influenza virus strains (such as A/VN/1203/04), can cause severe and often fatal infections. The field has used DNA microarrays and functional analyses to define the virus–host interactions that regulate influenza virus pathogenesis. These studies have identified gene expression signatures that correlate with virulence and have revealed that the timing and magnitude of the host response is a crucial determinant of the eventual outcome of infection.

This phenomenon was first demonstrated in studies that used a combination of mouse and macaque infection models and genome-wide transcriptional profiling to measure the host response to the reconstructed 1918 pandemic virus. In these animal models, the virus causes a rapidly fatal infection marked by severe lung pathology, intense neutrophil infiltration and the rapid and sustained induction of pro-inflammatory cytokine and chemokine genes, an event often referred to as a cytokine storm. By contrast, macaques infected with a highly pathogenic avian H5N1 influenza virus strain show a rapid and intense induction of interferon (IFN) and innate immune genes, but this eventually resolves as the animals recover. Genomic analyses have also revealed that macaques infected with the 1918 pandemic strain and mice infected with H5N1 influenza virus A/VN/1203/04 show a strong induction of genes encoding inflammasome components (for example, NLRP3 (NOD-, LRR- and pyrin domain-containing 3) and interleukin-1β) in this H5N1 virus is particularly virulent in mice, and although the inflammasome is part of the innate immune response to influenza A viruses,
the excessive activation of this response seems to be detrimental in this host.

As high-throughput data have accumulated in public databases, it has become possible to use this information to carry out meta-analyses. This strategy has been used to analyse data from a compendium of published studies that used mouse models to measure host transcriptional responses to lethal or non-lethal strains of influenza virus, respiratory syncytial virus or severe acute respiratory syndrome coronavirus (SARS CoV)\(^\text{96}\)\. Two alternative methods were used to generate gene expression signatures that are predictive of high or mild pathogenicity (defined here as 100% mortality and 100% survival, respectively). The first signature consists of 74 genes for which expression changes in the opposite direction in highly versus mildly pathogenic infections, with respect to mock infections (referred to as a ‘digital’ relationship). The second signature consists of 57 genes that are differentially expressed between highly and mildly pathogenic infections, without reference to mock infections (referred to as an ‘analogue’ relationship). Most genes in the analogue signature are differentially expressed during both lethal and non-lethal infections (compared with mock infections), but high pathogenicity corresponds with a higher degree of differential expression.

When the two signatures were tested for their ability to predict pathogenicity, the best predictor of a highly pathogenic infection was the analogue signature (FIG. 2). Significantly, this meta-analysis did not take into account the time after infection; that is, data from samples isolated at 1 day and at 5 days post-infection were treated equally. However, the majority of samples that were correctly identified as being from either a mildly or a highly pathogenic infection were from early or late time points, respectively. It is therefore likely that taking time post-infection into account would yield an even more accurate signature.

From these studies, it is now apparent that highly pathogenic respiratory viruses induce or suppress the expression of many of the same genes as mildly pathogenic viruses, but to a greater degree (and with different kinetics). Therefore, knowing the identity of genes that are differentially expressed in response to infection provides only part of the information needed to predict pathogenicity. The magnitude and timing of the host response are crucial determinants of the eventual disease outcome and might have important implications for antiviral therapy. To date, efforts to target the host response with a range of anti-inflammatory drugs have been largely unsuccessful\(^\text{97}\)\, and it is likely that effective host-directed therapy will depend not only on the target, but also on exactly when elements of the host response are suppressed or enhanced. Moreover, these findings point to the need for computational approaches that can describe nonlinear relationships (see below) and account for various factors associated with large multivariate data sets\(^\text{21}\)\. This also suggests the need for a paradigm shift in biomarker discovery to one that looks at sets of quantitative molecular measurements.

**Signatures of vaccine efficacy.** The application of systems-level analyses to vaccine research — variously termed systems vaccinology\(^\text{22}\) or vaccinomics\(^\text{23}\) — has led to the identification of molecular signatures that are predictive of vaccine immunogenicity and to new insights into the mechanisms of action of vaccines. In one of the first large-scale uses of this strategy, gene expression profiling and computational methods were used to identify gene expression signatures predictive of the strength of human adaptive immune response to the yellow fever vaccine, YF-17D\(^\text{24}\)\.

Transcriptional profiling of peripheral blood mononuclear cells from vaccinated subjects revealed that YF-17D induces the expression of genes encoding proteins that are associated with viral recognition and with transcription factors that regulate type I IFNs. Although also characteristic of the transcriptional response to active viral infection, this response did not correlate with subsequent CD8\(^+\) T cell or neutralizing-antibody responses, which are thought to mediate protection. However, an alternative computational and classification method, discriminant analysis via mixed integer programming (DAMIP), identified a signature (consisting of complement system and stress response genes) that is highly accurate in predicting subsequent CD8\(^+\) T cell activation. This method also identified a separate signature (which includes tumour necrosis factor (TNF) receptor superfamily members) that accurately predicts neutralizing-antibody expression during infection. The robustness of these signatures was verified through the analysis of samples from a different group of subjects vaccinated with a different lot of vaccine, thereby identifying new correlates of vaccine immunogenicity\(^\text{24}\)\.

More recently, similar systems approaches have been used to evaluate innate and adaptive immune responses to vaccination against influenza viruses, with the goal of identifying early gene expression signatures that correlate with immunogenicity\(^\text{25}\). Over a 3-year period, a
A series of clinical studies was undertaken in which young adults were vaccinated with either inactivated influenza vaccine (TIV) or live-attenuated influenza vaccine (LAIV). Molecular signatures for predicting antibody responses were identified by combining gene expression profiling, antibody response data, and DAMIP. The resulting predictive signature consisted of genes with known roles in antibody responses and other genes with previously unidentified roles in antibody or B cell responses. For example, one gene from the predictive signature, Camk4, encodes calcium–calmodulin-dependent protein kinase type IV (CaMKIV), a protein known to be involved in multiple immune system processes. However, it was not known whether this protein had a role in antibody responses. To demonstrate the ability of the systems biology approaches used in this study to identify biologically significant targets, Camk4-knockout and wild-type mice were vaccinated with TIV, and the Camk4-knockout mice exhibited significantly higher antibody titres than wild-type mice 7, 14 and 28 days after vaccination, thus revealing that Camk4 is important in regulating B cell responses. Although further investigations are needed to confirm many of these signature predictions, these studies demonstrate that systems approaches can both identify biological targets and generate new testable hypotheses related to the mechanism of vaccine action.

**An expanding view of the transcriptome.** Until recently, transcriptional profiling depended on the use of microarrays to measure the expression of well-annotated protein-coding genes. The advent of next-generation sequencing (BOX 2), however, has brought the ability to rapidly sequence the entire RNA complement of cells or tissues. This has led to a much expanded concept of the host transcriptome, as most recently revealed by the Encyclopedia of DNA Elements (ENCODE) project. It is now apparent that as much as three-quarters of the human genome is capable of being transcribed and that cells contain vast numbers and varieties of non-protein-coding RNAs. Some of these non-coding RNAs, such as microRNAs, have been well studied and are known to have roles in viral infection. For most others, functionality is less clear, but there is growing evidence that long non-coding RNAs also play a part in transcriptional and epigenetic gene regulation and in disease. RNA sequencing (RNA-seq) analysis of the host response to SARS CoV infection has revealed the differential expression of a range of host long non-coding RNAs (>200 nucleotides long) in lung samples from virus-infected mice. Many of these RNAs have similar expression patterns in vitro during influenza virus infection and during type I IFN treatment, suggesting that they are involved in regulating the innate immune response to a range of viruses. Expanding these analyses to include the sequencing of small RNAs also...
Box 2 | Next-generation sequencing

Next-generation sequencing is a massively parallel sequencing-by-synthesis approach that has replaced the first-generation automated Sanger sequencing method for genome analysis. The approach generates hundreds of millions of bases of high-quality DNA sequence per run and has enabled the rapid sequencing of entire genomes. It is also used for transcriptome profiling, an approach referred to as RNA sequencing (RNA-seq). This has a number of advantages over microarray-based profiling, including the ability to identify and quantify rare or as-yet-undiscovered transcripts and the ability to provide information on alternative splicing, genetic variation, and gene and exon boundaries. The method used for the construction of the CDNA library determines the type of transcripts to be sequenced; these include small RNAs (<200 bases; for example, microRNAs, small nucleolar RNAs and PIWI-interacting RNAs), mRNAs (various types of polyadenylated transcripts) or whole transcriptomes (polyadenylated and non-polyadenylated transcripts). After sequence reads are generated, they are aligned to a known reference genome, or if no reference genome is available, sequences can be assembled de novo. Because of the short length of sequence reads, the alignment and assembly of sequence information is a significant computational challenge. Additional potential applications of next-generation sequencing that are relevant to systems virology include the detection of binding events involving protein–DNA (ChIP–seq) or protein–RNA ( RIP–seq and CLIP–seq) (crosslinking immunoprecipitation followed by high-throughput sequencing)) and DNA methylation profiling (Methyl–seq)..

Given the advantages of next-generation sequencing, some have suggested that it will soon replace the use of microarrays for transcriptomic profiling. However, the enormous amount of data generated by next-generation sequencing requires an extensive information technology and computational infrastructure for data processing, storage and analysis. These costs and complexities, coupled with the cost of the sequencing itself, make it likely that microarrays will continue to be widely used for some time.

Small nucleolar RNAs (snoRNAs). RNAs that guide the modification (for example, methylation or pseudouridylation) of other RNAs, particularly ribosomal RNAs.

PIWI-interacting RNAs (piRNAs). Small RNAs that are thought to be involved in gene silencing through the formation of ribonucleoprotein complexes with PIWI proteins.

RIP–seq. Immunoprecipitation of RNA-binding proteins followed by high-throughput sequencing of the bound RNA.

CLIP–seq. (Crosslinking immunoprecipitation followed by high-throughput sequencing). A screening method used to identify RNA sequences that interact with either RNA-binding proteins or other RNAs.

revealed the differential expression of more than 200 small RNAs, such as small nucleolar RNAs (snoRNAs) and PIWI-interacting RNAs (piRNAs), in response to SARS CoV and influenza virus infections.

Similarly, RNA-seq has revealed that an HIV-infected CD4+ T cell line exhibits its differential expression of host microRNAs, snoRNAs and pseudogenes compared with uninfected cells.

Viral mRNA constitutes a surprisingly large portion of the total RNA in HIV-infected CD4+ T cells (in this study, nearly 40% by 24 hours after infection), and reads mapping to the viral genome have revealed novel viral RNA splice variants. A correlative analysis that combined mRNA-seq and small RNA-seq data suggested additional roles for host microRNAs in T cell activation and in transcriptional and cell cycle regulation during HIV infection.

Together, these studies attest to the power of RNA-seq to provide entirely new views of the transcriptional landscape and to highlight the previously unanticipated changes in transcription that occur in response to viral infection. Such insights are not limited to host transcription, as a combination of RNA-seq and mass spectrometry recently revealed that human cytomegalovirus (HCMV; a 240 kb DNA virus) produces hundreds of previously unidentified transcripts and short proteins that might have functional, regulatory or antigenic properties. Although the functional significance of changes in non-coding-RNA expression is only beginning to be examined, a better grasp of non-coding-RNA expression and function will certainly be necessary for a complete understanding of viral pathogenesis and the innate and adaptive immune responses, and for a more general view of gene regulation in the context of viral infection. Unfortunately, despite the advances that can be made using RNA-seq, the extensive computing infrastructure needed to handle large data files and the computational prowess required to align, assemble and analyse short sequence reads continues to put the approach out of reach for most laboratories.

Beyond the transcriptome

Of course, gene expression profiling provides only one measure of the host response to infection. In recent years, advances in systems-wide technologies have facilitated a ‘multi-omics’ approach that includes proteomics, metabolomics, lipidomics and elucidating virus–host protein interactions. All these measurements are adding to our understanding of the host response to viral infection, and new abilities to evaluate the role of host genetics and epigenetics are adding additional layers of complexity.

Alterations in cellular metabolism. Viruses have long been known to cause changes in host metabolism; however, the full extent of such changes was not clear until systems approaches were used to evaluate the metabolic reprogramming that results from HCMV infection. Using liquid chromatography–mass spectrometry to directly measure the levels of more than 160 different metabolites, it was discovered that HCMV infection induces large increases in numerous metabolites, including glycolytic and tricarboxylic acid (TCA) cycle intermediates, amino acids, NADH and pyrimidine. The metabolic signature induced by HCMV infection is readily distinguishable from the signature associated with the transition of quiescent cells into the G1 phase of the cell cycle, revealing the replacement of cellular metabolic homeostasis with an HCMV-specific metabolic programme. Systems-level metabolic flux profiling produced a first-of-a-kind metabolic map showing linkages between compounds and quantitative information about metabolic activity. The map is biochemically revealing, indicating that there is a global upregulation of metabolism by HCMV, with the greatest increase in the TCA cycle and its efflux to feed fatty acid biosynthesis.

To determine whether this reprogramming of the host metabolome is cell type or virus specific, fibroblast and epithelial cells were used to compare the host response to two strains of HCMV and two strains of herpes simplex virus type 1 (HSV-1). All four viruses produced significant changes in approximately 50% of the metabolome. Interestingly, the changes are consistent across different strains of the same virus and across cell types, but differ markedly between HCMV and HSV-1, demonstrating that these viruses induce distinct metabolic programmes. The findings derived from these systems-level approaches generated the hypotheses needed to drive additional focused studies in which more traditional methods were used to investigate the molecular mechanisms underlying the virus-specific hijacking of the host metabolome and the relevance of these mechanisms to potential therapeutic interventions.
Genomic and lipidomic analyses have also revealed that infection of primary bone marrow-derived macrophages with mouse cytomegalovirus results in a downregulation of metabolites involved in the cholesterol metabolism pathway. The lowering of cholesterol levels is mediated through the IFN-dependent downregulation of sterol regulatory element-binding protein 2 (SREBP2), a transcription factor that regulates sterol biosynthesis. Pharmacological or RNAi-mediated inhibition of the sterol pathway was shown to result in increased protection against viral infection in cell culture and in mice, demonstrating the potential benefit of targeting a host metabolic pathway as an antiviral strategy.

**Virus–host interactomes.** Based on the premise that viral proteins interact with cellular factors to promote efficient viral replication and pathogenesis, systems-wide small interfering RNA (siRNA) or short hairpin RNA (shRNA) screens, yeast two-hybrid libraries and bioinformatic methods are being used to construct and describe virus–host interactomes and in turn identify cellular targets for therapeutic intervention. Such interactomes have been generated for numerous viruses, including influenza virus, HIV, dengue virus, hepatitis C virus (HCV), herpesviruses and SARS CoV, and have yielded ‘hit lists’ of cellular factors that might be important in viral pathogenesis.

Although the number of interactions identified by these studies is impressive, few of the genes identified have been subjected to functional analyses to confirm their role in viral replication. Moreover, there is little overlap in the host factors identified by the different screens. This could be due to a variety of factors, including variation in the screening systems, cell types and viruses used, as well as differences in the methods applied to identify interacting partners. The benefits and shortcomings of these studies, along with factors affecting their outcome, have been the subject of several detailed reviews.

To fully realize the collective information residing in this vast collection of data, it will be necessary to develop computational and mathematical methods which are capable of fully integrating interactomes (as well as their associated metadata) that have been constructed using different methods or obtained from different biological systems. In a first attempt at such integration, meta-analysis of virus–host interactome data for five distinct viruses identified both common and virus-specific human protein targets. Common targets include proteins involved in the cell cycle, apoptosis, the unfolded-protein response and nuclear transport. Many of the common host targets identified are multifunctional hubs — that is, they have multiple functions or roles within the cell. When coupled with the fact that many viral proteins interact with multiple host proteins, this combination of factors might explain how viruses, with their relatively small genomes, are capable of dysregulating so many aspects of host biology. Unfortunately, because these common host targets often have numerous cellular functions, focusing on them as drug targets might be problematic.

In an interesting twist on these studies, virus–host interactome data, together with data on host transcriptional changes resulting from the expression of 123 viral ORFs derived from DNA tumour viruses, were used to predict cellular genomic variations (for example, mutations, deletions or translocations) that can give rise to cancer. By defining the rewiring of cellular networks and pathways caused by the viral proteins and identifying a list of host proteins that are central to the rewiring, the systematic identification of host targets of DNA tumour viruses was found to be as successful as traditional large-scale cataloguing of tumour mutations for cancer gene identification. This suggests that disease phenotypes, whether resulting from viral infection or cancer, might be the result of network perturbations rather than individual genetic or genomic variations.

**Host genetics and epigenetics.** Host genetic variation has typically been thought of as a confounding factor that limits the ability to draw conclusions from data obtained using outbred (for example, human and nonhuman primate) populations. More recently, attempts are being made to better understand how genetic diversity influences infection outcome and how knowledge of genetic diversity can be incorporated into the construction of robust and predictive network models. One particularly exciting example is the Collaborative Cross mouse resource. The Collaborative Cross is a unique panel of multiparental recombinant inbred mouse strains designed to capture the level of genetic diversity found in outbred populations; this provides a resource for systematically identifying individual and multiple host genetic traits that contribute to complex immune phenotypes and disease outcome (BOX 3).

As an example, a genetically diverse panel of pre- Collaborative Cross mice (not fully inbred) determined to have severe or mild responses to influenza virus infection was used to identify expression quantitative trait loci (eQTLs) associated with the host response to infection. Twenty-one high-confidence eQTLs were identified, 17 of which were confirmed using mice from the eight Collaborative Cross founder strains. Many of these genes have known functions related to immunity or the host response to infection, such as Ifi27l2a (IFNα-inducible 27-like 2A), Clec16a, Pde7a (high-affinity cyclic AMP-specific 3′,5′-cyclic phosphodiesterase 7A) and Tcf7l1 (transcription factor 7-like 1), whereas for others, such as the serine/threonine protein kinase gene Sik1 and sentrin-specific protease 5 (Semp5), their role in influenza virus infection is not yet clear. Structural equation modelling was used to identify potential regulatory relationships between additional genes and the validated eQTLs, suggesting that these genes and corresponding subnetworks have important roles in the host response by either promoting a protective (mild) or a pathologic (severe) response to influenza virus infection, depending on the specific genetics of the host. Thus, using this genetically diverse population, high-confidence gene candidates involved in regulating the host response to influenza virus infection were identified, allowing for future investigations into their utility as therapeutic targets.
The Collaborative Cross is a genetic reference population of recombinant inbred (RI) mice derived from eight laboratory mouse strains. Such populations are used to study complex traits and are most beneficial when they contain large numbers of lines that exhibit substantial and uniform genetic variation. In the case of the Collaborative Cross, five inbred strains (A/J, C57BL/6J, 129S1/SvImJ, NOD/ShiLtJ and NZO/HllJ) and three wild-derived strains (CAST/EiJ, PWK/PhH and WSB/EiJ) are being used to generate a large panel of RI lines through a randomized funnel breeding scheme, resulting in hundreds of independent RI lines in which genetic variation is randomly and uniformly distributed throughout the genome. When they are established, RI lines can be intercrossed among themselves to generate recombinant intercross (RIX) lines or backcrossed to other non-Collaborative Cross lines to generate RI backcross (RIB) lines. RIX lines maximize genetic diversity, and because they are outbred, they are better models of human populations and can provide information on gene dominance and epistasis. RIB lines can be used, for example, to evaluate whether specific allele combinations alter the phenotype associated with a given target gene (such as a dominant targeted gene knockout). For example, a comparison between viral titre in the lungs and the extent of airway inflammation for a panel of pre-Collaborative Cross (not fully inbred) mice demonstrates the impact of genetic diversity on influenza virus disease phenotypes (see the figure; each diamond represents a different mouse). Phenotypes range from high viral titre and high inflammation to low viral titre and little inflammation; in addition, some animals show unique phenotypes, such as a high viral titre with little accompanying inflammation. The Collaborative Cross is being developed as a resource for the biomedical research community, and genotype information and mouse line availability is publically available at University of North Carolina (UNC) Systems Genetics, Chapel Hill, USA.

**Box 3 | The Collaborative Cross**

The Collaborative Cross is a genetic reference population of recombinant inbred (RI) mice derived from eight laboratory mouse strains. Such populations are used to study complex traits and are most beneficial when they contain large numbers of lines that exhibit substantial and uniform genetic variation. In the case of the Collaborative Cross, five inbred strains (A/J, C57BL/6J, 129S1/SvImJ, NOD/ShiLtJ and NZO/HllJ) and three wild-derived strains (CAST/EiJ, PWK/PhH and WSB/EiJ) are being used to generate a large panel of RI lines through a randomized funnel breeding scheme, resulting in hundreds of independent RI lines in which genetic variation is randomly and uniformly distributed throughout the genome. When they are established, RI lines can be intercrossed among themselves to generate recombinant intercross (RIX) lines or backcrossed to other non-Collaborative Cross lines to generate RI backcross (RIB) lines. RIX lines maximize genetic diversity, and because they are outbred, they are better models of human populations and can provide information on gene dominance and epistasis. RIB lines can be used, for example, to evaluate whether specific allele combinations alter the phenotype associated with a given target gene (such as a dominant targeted gene knockout). For example, a comparison between viral titre in the lungs and the extent of airway inflammation for a panel of pre-Collaborative Cross (not fully inbred) mice demonstrates the impact of genetic diversity on influenza virus disease phenotypes (see the figure; each diamond represents a different mouse). Phenotypes range from high viral titre and high inflammation to low viral titre and little inflammation; in addition, some animals show unique phenotypes, such as a high viral titre with little accompanying inflammation. The Collaborative Cross is being developed as a resource for the biomedical research community, and genotype information and mouse line availability is publically available at University of North Carolina (UNC) Systems Genetics, Chapel Hill, USA.
of the interaction networks between host proteins and HCV core and NS4B proteins has been used to identify host proteins that constitute potential anti-HCV therapeutic targets, including α-enolase, paxillin and a solute carrier protein (SLC25A5). A better understanding of network topology will not only provide the opportunity to identify potential targets for therapeutic intervention, but also offer insights into possible off-target effects on network signalling that might be induced by drug treatment.

Although much can be learned from the construction and topological analysis of host–pathogen interaction networks using samples from whole tissues, the heterogeneity of cell types present in most tissues means that such networks provide a generalized picture of the changes that occur in the host during the course of infection. For example, it is difficult to delineate from these types of networks the signalling events that might occur between infected lung epithelial cells and cells of the immune system, both inside and outside the infected tissue. These intracellular interactions are also controlled by signal transduction pathways, which communicate signals from the extracellular environment to intracellular effector processes. To fully interpret the data, we need a much better understanding of intracellular signalling processes, the cells that are involved and the directionality of their effects on infection outcome.

A few studies have begun to explore cell type-specific and intracellular signalling on a system-wide scale. For example, flow cytometry and gene expression data from bronchoalveolar lavage (BAL) fluid from young adult and aged macaques infected with 2009 pandemic H1N1 influenza virus were analysed in conjunction with data from the Immune Response In Silico (IRIS) database. This database contains cell type-specific gene expression patterns associated with various types of immune cells. By computationally comparing differentially expressed genes in BAL with cell type-specific gene expression patterns in the IRIS database, it was possible to identify genes associated with specific immune cell types, including activated dendritic cells, CD4+ and CD8+ T cells and naive B cells. In particular, genes associated with B cell and T cell markers were more highly upregulated in young adult animals.

**Figure 3 | Co-regulation networks.** a | A schematic showing one bottleneck gene and two hub genes in a hypothetical network. Hubs are highly connected to other nodes in the network. Bottlenecks have a high degree of betweenness centrality and connect or act as bridges between subnetworks. b | The context likelihood of relatedness method was used to infer functional associations between differentially expressed genes responding to H5N1 influenza virus infection in mice. Genes (nodes) are coloured according to the ratio of gene expression values for high-dose infection to low-dose infection. Red indicates higher expression and blue indicates lower expression in the high dose than in the low dose. The complexity of these networks does not allow a visual determination of bottlenecks. Rather, the betweenness centrality is calculated for each node, and the size of the node is relative to this value. The larger nodes are major bridges between different parts of the network. c | A major hub from a separate subnetwork within the H5N1 influenza virus co-regulatory network.
Recent studies of mouse models of intestinal inflammation induced by TNF treatment provide another good example of how systems approaches can be used to evaluate signalling between cell types in a complex tissue environment. In these studies, the authors combined flow cytometry measurements and phosphoprotein, cytokine and chemokine expression data from various immune cell types over time and under diverse conditions and were thus able to construct statistically robust multivariate regression models that related the phosphoprotein signals, cytokines, chemokines and cell types to specific phenotypes. These models helped to elucidate key molecular and cellular processes governing epithelial cell apoptosis and proliferation in response to TNF treatment. For example, monocyte chemotactic protein 1 (MCP1; also known as CCL2) was predicted by the model to be especially protective against TNF-induced apoptosis, and this hypothesis was confirmed by treating with an MCP1-specific antibody before TNF administration. The model also indicated that plasmacytoid dendritic cells might be particularly important for inducing apoptosis, and depleting these cells from mice under conditions that normally produced the most severe epithelial cell apoptosis did indeed revert the TNF-induced phenotype to the mildest outcome.

In addition to constructing network models that span intra- and intercellular signalling processes, it will be necessary to consider nonlinear relationships such as how the network functions over time (that is, the dynamics of the system). This is particularly true in light of evidence (discussed above) that the magnitude and timing of the host response to respiratory viruses are crucial determinants of the eventual disease outcome. Similar evidence is accumulating that the outcome of HIV infection is also related to the activation dynamics of host gene regulation. However, high-throughput data are typically static and often not adequate for modelling dynamic systems. To help overcome this limitation in analysing network dynamics, inference methodologies are being devised to reinterpret activity differences caused by system perturbations as differences in observation time. For example, changes in pathogen-induced gene expression that are associated with genetic variability, whether in the pathogen (for example, mutant viruses) or the host (as occurs in the Collaborative Cross mouse model), can potentially be used to indirectly infer the crucial dynamics of a system without having to measure the system over time. This innovative approach has only recently become feasible with major breakthroughs in the theory of dynamic systems and in geometrical high-dimensional analysis methodology.

As virology continues to transition into a more quantitative science, increasing attention must be paid not only to network dynamics but also to other nonlinear interactions, such as cooperative or synergistic relationships, which characterize so much of biology. Current biomarker discovery and the identification of molecular predictors of adjuvant, vaccine or, more generally, drug efficacy are largely unsuccessful because nonlinear interactions between molecules, as well as genetic diversity in populations, are not taken into sufficient consideration. Geometrical methods (that is, methods used to identify the structure in data by identifying spatial and temporal relationships) are increasingly being used in the analysis of high-throughput molecular data. In particular, novel combinations of geometrical methods, such as those based on singular value decomposition (SVD) and multidimensional scaling (MDS), are beginning to be used in systems virology to better understand the nonlinear interactions between variables and isolate those interactions from biological noise.

Liver biopsies were carried out on patients who were infected with HCV and had received liver transplants, and SVD–MDS analysis of transcriptomic data from these biopsy samples, in combination with categorical analysis (to take into account variables such as age, time post-transplantation and fibrosis score), has been used to identify a molecular signature for patients at risk of developing severe fibrosis. SVD–MDS and co-abundance networks (which relate molecules to each other on the basis of their abundance profiles) were also used to integrate proteomic and metabolomic data sets obtained from the same cohort of patients. This strategy identified a potential role for oxidative stress in rapid fibrosis progression after the transplant surgery and identified serum metabolites that might prove useful as biomarkers for predicting progression to fibrosis. This understanding of network structure can now be used to simulate human liver metabolism using novel flux balance modelling approaches to better comprehend and eventually treat disease. Additional geometrical approaches have been described and should prove useful for effectively bridging different technologies and for integrating diverse types of data, thereby enabling better analysis of the data that are already available in public databases and repositories. Finally, geometrical methods, as well as links between geometry, information theory and probability theory, will also help to identify causal relationships, which is clearly a currently unmet challenge. Unlike purely statistical approaches, geometrical methods can be used to integrate different individual measures for the purposes of comparison and combination into coherent objects that identify relationships between genes, transcripts or proteins.

Conclusions

After having been extensively hyped as a paradigm shift, systems-level approaches have since been criticized for failing to rapidly fulfil their initial grand promises. Standing in the way have been numerous technical, experimental and mathematical hurdles. However, as discussed in this Review, significant progress is being made, and new computational approaches are leading the way. Also of importance is the ever-growing availability of high-throughput data in public databases; indeed, data sharing is crucial to the future success of systems virology. To date, bench work has been the necessary precedent to computational approaches, but we are now at a point where there are sufficient data available that computational methods can be the starting point for making discoveries, generating hypotheses and, in turn, guiding targeted bench work. A prime
Box 4 | The importance of data sharing

The rapid dissemination of high-throughput omics data sets to the scientific community is an important element for the future success of systems virology. The complexity of virus–host interactions can be unravelled only through meta-analyses that rely on advanced computational and mathematical approaches and the integration of large, coherent and systematic multi-dimensional data sets. Owing to the immense amounts of time, expertise, manpower and expense that are required to produce these data sets, it is imperative for the systems virology community to share data in a timely manner. In most cases, owing to the richness of information contained in these data sets, this can be done with little chance of compromising one’s own research. Two of the US National Institute of Allergy and Infectious Diseases (NIAID)-sponsored Bioinformatics Resource Centers, ViPR (Virus Pathogen Resource) and IRD (Influenza Research Database), are charged with capturing, publicly sharing, storing, integrating and visualising systems-wide high-throughput omics data sets that detail the host response to viral infection. Importantly, in addition to capturing data types such as transcriptomic, proteomic, metabolomic, lipidomic and ChIP-seq (chromatin immunoprecipitation followed by sequencing) data, the goal of these databases is to capture extensive metadata that provide users with the experimental details needed to facilitate data interpretation. The practice of linking metadata to several forms of the omics data (be it raw, analysed or modelled) enables ViPR and IRD to meet the demands of a range of end users (see the figure). For instance, at their portals, users can search for information on individual genes, download lists of differentially expressed genes or link metadata to raw data for doing meta-analysis. All laboratories, as a service to the community, should be committed to sharing these types of data sets, whether through organization-specific public websites, the appropriate NIAID Bioinformatics Resource Center or a combination of both.

In conclusion, contemporary virology cannot afford to simply catalogue myriad circumstantial observations. Systems-level approaches provide the opportunity to assemble the incomplete puzzle of biology in a meaningful way that will advance our understanding of how viruses cause disease and lead to improved patient care. Although there will always be a need for traditional microbiology, the success of undergraduates in this day and age will depend on their ability to combine traditional skills with systems approaches and mathematics. The time is ripe, the data are here, and the mathematics to put them together is coming along. Take heed of the words on Plato’s doorstep: “γνωμένης μηδένς εισέλθι” (“Let no one ignorant of geometry enter”).

1. Aderem, A. et al. A systems biology approach to infectious disease research: innovating the pathogen-host research paradigm. mBio 2, e00525–10 (2011).
2. Tisoncik, J. R. & Katze, M. G. What is systems biology? Future Microbiol. 5, 139–141 (2010).
3. Katze, M. G. et al. Systems Biology (Springer, 2015).
4. Macluan, C. Systems biology: evolving into the mainstream. Cell 144, 859–861 (2011).
5. Diercks, A. & Aderem, A. Systems approaches to dissecting immunity. Curr. Top. Microbiol. Immunol. 365, 1–19 (2013).
6. Lauffenburger, D. A. The multiple dimensions of integrative biology. Integr. Biol. (Camb.) 4, 9 (2012).
7. Lesne, A. Multiscale analysis of biological systems. Acta Biotheor. 61, 3–19 (2013).
8. Geiss, G. K. et al. Large-scale monitoring of host cell gene expression during HIV-1 infection using cDNA microarrays. Virology 266, 8–16 (2000).
9. Fukushima, S. & Kanawi, Y. The pathogenesis of influenza virus infections: the contributions of virus and host factors. Curr. Opin. Immunol. 23, 481–486 (2011).
10. Korth, M. J., Tichy, N., Benecke, A. G. & Katze, M. G. Systems approaches to influenza-virus host interactions and the pathogenesis of highly virulent and pandemic viruses. Semin. Immunol. 4 Dec 2012 (doi:10.1016/j.smim.2012.11.001).
11. Kash, J. C. et al. Genomic analysis of increased host immune and cell death responses induced by 1918 influenza virus. Nature 445, 578–581 (2006).
12. Kobasa, D. et al. Aberrant innate immune response in lethal infection of macaques with the 1918 influenza virus. Nature 445, 319–325 (2007). The first study in which non-human primates are infected with the 1918 pandemic influenza virus. This study highlights the importance of the ability of the virus to modulate the host response.
13. Tisoncik, J. R. et al. Into the eye of the cytokine storm. Microbiol. Mol. Biol. Rev. 76, 16–52 (2012). A review discussing the cytokine storm in the context of viral infections, and how systems virology approaches have provided significant insights into the kinetics of cytokine gene expression.
14. Baskin, C. R. et al. Early and sustained innate immune response defines pathology and death in nonhuman primates infected by highly pathogenic influenza virus. Proc. Natl Acad. Sci. USA 106, 5455–5460 (2009).
In the context of viral infection, the viral genome is translated into viral proteins that are required for the virus life cycle. These viral proteins can interact with host proteins, leading to the modulation of host cellular processes. For instance, within the context of the human cytomegalovirus (HCMV) infection of human embryonic kidney (HEK) 293 cells, our study revealed that viral proteins can alter the expression patterns of host genes identified using microarray analysis. This alteration in gene expression can lead to the modulation of host cellular processes and the initiation of cellular responses such as apoptosis or inflammation.

In this study, we performed a systems-level analysis of host gene expression in response to HCMV infection. Our results showed that the expression of several genes was significantly altered, including those involved in the regulation of immune responses, inflammation, and cell cycle. These findings suggest that HCMV infection can alter the host cellular environment, leading to the initiation of immune responses and the modulation of cellular functions.

Furthermore, our study also revealed that the expression of certain host genes was downregulated by HCMV infection. These genes were involved in metabolic pathways, including lipid metabolism. The downregulation of these genes by HCMV infection may contribute to the pathogenesis of HCMV infection, as it can lead to the alteration of the host metabolic state, which can facilitate viral replication.

In conclusion, our study provides insight into the host cellular responses to HCMV infection and highlights the importance of host gene expression in the modulation of the viral life cycle. Understanding these responses can lead to the development of new strategies to combat HCMV infection and prevent the spread of this infectious disease.
REVIEWS

84. Diamond, D. L. et al. Proteome and computational analyses reveal new insights into the mechanisms of hepatitis C virus-mediated liver disease posttransplantation. Hepatology 56, 28–38 (2012).
85. Sonnenschein, N. et al. A network perspective on metabolic inconsistency. BMC Syst. Biol. 6, 41 (2012).
86. Poonapalli, S. P., Saunders, M. A., Van Loan, C. F. & Alter, O. A higher-order generalized singular value decomposition for comparison of global mRNA expression from multiple organisms. PLoS ONE 6, e28072 (2011).
87. Huang, S. S. & Frankel, E. Integrating proteomic, transcriptional, and interactome data reveals hidden components of signaling and regulatory networks. Sci. Signal. 2, ra40 (2009).
88. Luksza, M., Lassig, M. & Berg, J. Significance analysis of a computational method (based on the Steiner tree problem) that provides a general framework for building models of regulatory networks from high-throughput data sets.
89. Josset, L. et al. Gene expression signature-based screening identifies new broadly effective influenza A antivirals. PLoS ONE 5, e11569 (2010).
90. Yang, W.-L. R., Lee, Y.-E., Chen, M.-H., Chao, K.-M. & Huang, C.-Y. In-silico drug screening and potential target identification for hepatocellular carcinoma using Support Vector Machines based on drug screening result. Gene 518, 201–208 (2015).
91. Hood, L. & Friend, S. H. Predictive, personalized, preventive, participatory (P4) cancer medicine. Nature Rev. Clin. Oncol. 8, 184–187 (2011).
92. Bengoechea, J. A. Infection systems biology: from reactive to proactive (P4) medicine. Int. Microbiol. 15, 55–60 (2012).
93. Southern, E. M. DNA microarrays. History and overview. Methods Mol. Biol. 170, 1–15 (2001).
94. Goren, F., Carbone, W. & Barazangi, D. in Molecular Profiling (eds Espina, V. & Liotta, L. A.) 89–105 (Humana, 2012).
95. Li, W. & Ruan, K. MicroRNA detection by microarray. Anal. Bioanal. Chem. 394, 1117–1124 (2009).
96. Fouse, S. D., Nagarajan, R. P. & Costello, J. J. Genome-scale DNA methylation analysis. Epigenomics 2, 105–117 (2010).
97. Lam, C.-W., Lau, K.-C. & Tong, S.-F. Microarrays for personalized genomic medicine. Adv. Clin. Chem. 52, 1–18 (2010).
98. Metzker, M. L. Sequencing technologies — the next generation. Nature Rev. Genet. 11, 31–46 (2010).
99. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. Nature Rev. Genet. 10, 57–65 (2009).
100. Martin, J. A. & Wang, Z. Next-generation transcriptome assembly. Nature Rev. Genet. 12, 671–682 (2011).
101. Park, P. J. ChIP-seq: advantages and challenges of a maturing technology. Nature Rev. Genet. 10, 669–680 (2009).
102. König, J., Zarnack, K., Luscombe, N. M. & Ule, J. Protein–RNA interactions: new genomic technologies and perspectives. Nature Rev. Genet. 13, 77–83 (2012).
103. Collaborative-Cross-Consortium. The genome architecture of the Collaborative Cross mouse genetic reference population. Genetics 190, 389–401 (2012).
104. Zou, F. et al. Quantitative trait locus analysis using recombinant inbred intercrosses: theoretical and empirical considerations. Genetics 170, 1299–1311 (2005).

Acknowledgements
The authors thank L. Josset for generating the networks in figure 3, and M. Heise and M. Ferris for providing the data used in box 3. Research in the author’s laboratory is supported by Public Health Service grants R2400001172, R2400011157, P50DA015625, P51RR00166 and U54AI081680, and by federal funds from the US National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under contract HHSN272200800060C.

Competing interests statement
The authors declare no competing financial interests.

FURTHER INFORMATION
Michael G. Katze’s homepage: http://virology.washington.edu
Collaborative Cross: http://www.ccmbio.unc.edu/CCstatus
ENCODE: http://encodeproject.org/ENCODE
IRD: http://www.fludb.org
Systems Influenza Center: http://www.systemsinfluenza.org
Systems Virology Center: https://www.systemsvirology.org
VIPP: http://www.viprbrc.org

ALL LINKS ARE ACTIVE IN THE ONLINE PDF