

# Visualization of omics data for systems biology

Nils Gehlenborg<sup>1,2</sup>, Seán I O'Donoghue<sup>3</sup>, Nitin S Baliga<sup>4</sup>, Alexander Goesmann<sup>5</sup>, Matthew A Hibbs<sup>6</sup>, Hiroaki Kitano<sup>7–9</sup>, Oliver Kohlbacher<sup>10</sup>, Heiko Neuweber<sup>5</sup>, Reinhard Schneider<sup>3</sup>, Dan Tenenbaum<sup>4</sup> & Anne-Claude Gavin<sup>3</sup>

High-throughput studies of biological systems are rapidly accumulating a wealth of 'omics'-scale data. Visualization is a key aspect of both the analysis and understanding of these data, and users now have many visualization methods and tools to choose from. The challenge is to create clear, meaningful and integrated visualizations that give biological insight, without being overwhelmed by the intrinsic complexity of the data. In this review, we discuss how visualization tools are being used to help interpret protein interaction, gene expression and metabolic profile data, and we highlight emerging new directions.

Visualization has long been key in helping to understand biological systems, such as metabolism<sup>1</sup>, signaling<sup>2</sup> and the regulation of gene expression<sup>3</sup>. In recent years, the study of such systems has been profoundly influenced by the development of a wide range of high-throughput experimental methods (Box 1), resulting in a greatly increased volume of complex, interconnected data. Remarkably, in spite of these changes, and in spite of the development of new methods for visualizing and analyzing these data, we still use the same primary visual metaphor to communicate ideas about biological systems: namely, pathways (graphs that show overall changes in state) or, more generally, networks (graphs that do not necessarily show state changes).

As high-throughput experimental methods have become more routine, many more scientists are using network and pathway visualization to record and communicate their findings. There are now over 300 web resources<sup>4</sup> (see <http://pathguide.org/>) providing access to many thousands of pathways and networks that document millions of interactions between proteins, genes and small molecules.

There has been a corresponding increase in the development of visualization tools for systems biology

data<sup>5–7</sup>. These tools are very diverse, but they can be broadly divided into two partly overlapping categories, the first consisting of tools focused on automated methods for interpreting and exploring large biological networks (Table 1), and the second consisting of tools focused on assembly and curation of pathways (Table 2). Many of these tools are tightly integrated with public databases, thus allowing users to visualize and interpret their own data in the context of previous knowledge.

For users and developers of these visualization tools, one of the key challenges is how to benefit from the explosion in systems biology data without being overwhelmed by it—or, in practical terms, how to present the data at the right level of detail, in a cohesive, insightful manner. Clearly, the answers depend on context.

In this review, we first discuss the methods and tools now being used to visualize and analyze data sets from three main types of high-throughput experiments: namely, the investigation of protein-protein interactions, of gene expression profiles and of metabolic profiles. Such experiments are used to study cellular response to a wide variety of conditions—including drug exposure, disease states and specific genetic

<sup>1</sup>European Bioinformatics Institute, Cambridge, UK. <sup>2</sup>Graduate School of Life Sciences, University of Cambridge, Cambridge, UK. <sup>3</sup>European Molecular Biology Laboratory, Heidelberg, Germany. <sup>4</sup>Institute for Systems Biology, Seattle, Washington, USA. <sup>5</sup>CeBiTec, Bielefeld University, Bielefeld, Germany. <sup>6</sup>The Jackson Laboratory, Bar Harbor, Maine, USA. <sup>7</sup>Sony Computer Science Laboratories, Tokyo, Japan. <sup>8</sup>The Systems Biology Institute, Tokyo, Japan. <sup>9</sup>Okinawa Institute of Science and Technology, Okinawa, Japan. <sup>10</sup>University of Tübingen, Tübingen, Germany. Correspondence should be addressed to S.I.O. (sean.odonoghue@embl.de).

## BOX 1 KEY EXPERIMENTAL METHODS FOR SYSTEMS BIOLOGY

**Oligonucleotide microarrays.** The most widely used methods to monitor the expression levels of RNA transcripts in a biological sample are based on microarrays. They measure the hybridization of fluorescently labeled cDNA, synthesized from extracted mRNA, to known nucleotide sequences spotted on solid surfaces<sup>117</sup>. For all genes on the microarray, an expression value is derived from the fluorescence intensity of the hybridized RNAs. These expression values are typically unitless and have meaning only in the context of a reference measurement. Before further analysis takes place, the measurements must therefore be normalized to remove systematic biases and to make it possible to compare measurements from different samples.

Quality assessment is likewise essential for the validity of later analyses. This is typically performed with the help of (platform-dependent) quality scores at the level of both individual probes and entire arrays, complemented by diagnostic visualization tools that have been developed for this purpose<sup>118,119</sup>. Evaluating the quality of individual arrays is routinely done with spatial intensity distributions plots and plots of intensity ratio versus mean intensity (**Supplementary Fig. 1a**). Comparison of multiple arrays can be achieved with intensity box plots, which are a practical tool to detect outlier arrays that should be excluded from subsequent analysis (**Supplementary Fig. 1b**). Several tools that provide quality assessment visualizations are listed in **Supplementary Table 1**.

**RNA deep sequencing.** The most recent transcriptomics approaches are based on the deep sequencing of transcripts extracted from biological samples<sup>33</sup>. The resulting sequence reads—typically 30 to 400 base pairs long, depending on the DNA-sequencing technology used—are then commonly aligned to a reference genome and evaluated to determine their quality.

Tools for data processing and quality assessment typically provide diagnostic visualizations. Examples include the R/Bioconductor packages ShortRead<sup>120</sup> and edgeR<sup>121</sup>. The latter provides many functions that are analogous to those in the limma package<sup>122</sup> for transcriptomics data from microarrays. Reads aligned to a genome can also be visualized and evaluated with some of the more recent genome browsers that can handle short read data, such as the Integrative Genomics Viewer (<http://www.broadinstitute.org/igv/>). This and similar tools are discussed in the accompanying review by Nielsen *et al.*<sup>72</sup>.

**Mass spectrometry.** In mass spectrometry (MS) experiments, the compounds present in a sample are identified through the accurate measurements of their mass-to-charge ratios. MS has applications in many fields, including proteomics, metabolomics and interactome mapping.

In proteomic applications, typical MS data sets consist of lists of proteolytic peptides characterized by their mass-to-charge ratios (MS spectra, MS1). These peptides can be further fragmented and measurements of the resulting mass spectra (MS-MS spectra or tandem MS spectra, MS2) used to deduce their sequences. In some cases, complex samples must be fractionated and proteolytic peptides are separated using high performance liquid chromatography (LC) before MS analysis (LC-MS).

Several search engines have been developed to predict peptides and proteins through the comparison of experimentally measured spectra to theoretical spectra (predicted from sequence databases). Quality scores provide a measure of the reliability of a given protein or peptide identification<sup>123</sup>. For example, for Mascot<sup>124</sup>, the most broadly used algorithm, the score features the number of identified peptides (sequence coverage).

The overall quality of entire MS data sets is generally measured by the false discovery rate (FDR), which is the 'expected' proportion of incorrect assignments among the accepted assignments. The most popular approach to calculate FDR is based on the use of a target-decoy database<sup>123</sup>. Also, an array of visualization tools has been developed to evaluate the technical quality of the samples and of MS runs. For example, the overall distribution of peptides in an LC-MS map can be visualized with Pep3D<sup>125</sup> or TOPPView<sup>126</sup>, enabling the detection of possible biases, the presence of chemical contaminants or poor separations during the LC (**Supplementary Fig. 2**). Additionally, Pep3D can integrate quality scores for individual protein or peptide identifications generated by search engines into these maps (**Supplementary Fig. 3**). We list tools for mass spectrometry data visualization and evaluation in **Supplementary Table 1**.

In metabolomics applications, owing to large chemical diversity and variation in molecular composition of the analytes, various chromatographic systems, such as gas chromatography (GC), LC or electrochemistry (EC), are generally applied before MS. GC-MS is the most popular method for global metabolite profiling<sup>127</sup>. It can be complemented with LC-MS analysis to identify compounds that are not suitable for GC-MS analysis<sup>128</sup>. Similarly to the approaches developed for peptides, metabolites can be identified on the basis of their fragmentation patterns, for which mass spectral fingerprint libraries are being developed. Because the raw data are of the same kind as in proteomics mass spectrometry studies, very similar visualization methods are used to assess data quality (**Supplementary Table 1**).

**Nuclear magnetic resonance.** Nuclear magnetic resonance (NMR) is a common method in metabolomics and, in contrast to MS-based approaches, in most cases does not require analyte separation. NMR spectroscopy can provide detailed information on the molecular structure of compounds found in complex mixtures, and a wide range of small molecule metabolites in a sample can be detected simultaneously. Biofluids, cell and tissue extracts can be analyzed with minimal sample preparation through the use of <sup>1</sup>H NMR spectroscopy<sup>129</sup>. With the use of two-dimensional NMR spectra, the identification and reliable quantification of individual metabolites becomes feasible, which enables NMR-based metabolite profiling. Data processing and spectral deconvolution are challenging, and databases of NMR spectra of pure metabolites are not yet comprehensive, but they nonetheless do already help in the identification process<sup>130</sup>. Applications such as MetaboMiner<sup>131</sup> can be used for the semiautomated identification of metabolites in two-dimensional NMR spectra, supported by visualizations that allow the scientist to inspect the matches of peaks to reference spectra and assess match quality.

perturbations (for example, gene deletions, gene insertions and siRNA knockdowns). Often, these experiments produce new knowledge that is then either added to existing pathways or used to create new pathways. Thus, we end with a discussion of methods and tools for pathway editing.

### Protein interaction data

A range of experimental methods are at present being used for high-throughput studies of protein interactions<sup>8</sup>. For instance, in yeast, pairwise interactions have been studied on the genome scale using yeast two-hybrid screens or protein complementation assays, whereas the assembly of proteins within complexes has been systematically charted using tandem affinity purification coupled with mass spectrometry (TAP-MS) (Box 1 and Supplementary Figs. 2 and 3). Recent analyses have estimated that, in yeast, some 20,000 pairwise interactions may take place between the ~5,000 gene products<sup>9</sup>, and about 800 protein complexes may exist<sup>10</sup>. As a result of these and similar studies in other species, vast amounts of protein interaction data are accumulating in public databases<sup>11,12</sup> such as DIP<sup>13</sup>, HPRD<sup>14</sup> and IntAct<sup>15</sup>.

The size and complexity of these data sets can be daunting; hence, a common general strategy is to iteratively dissect the data sets into smaller subsets. Typically, these subsets are defined as sets of proteins that belong to the same complex, or that are found at the same subcellular location, or that belong to a similar functional category. Visualization is key in this strategy, as human judgment and intervention are often needed, in part because of errors (false positives and false negatives) in protein interaction data sets<sup>9,16</sup>. Many visualization tools have been developed specifically to support the analysis of protein networks (Table 1); here we discuss how these tools can be used to help dissect large data sets of interactions, extract biological insight and generate hypotheses leading to further experimental investigations.

As proteins rarely act alone, a first step in analyzing a protein interaction data set is to identify protein complexes and groups of complexes. For small, simple networks, visualized as a graph in which each node represents a protein and each edge represents an interaction between two proteins, the arrangement of proteins and complexes can usually be seen clearly using a standard 'force-directed' layout<sup>17</sup>, which automatically arranges each node

**Table 1** | Visualization tools focused on interaction networks

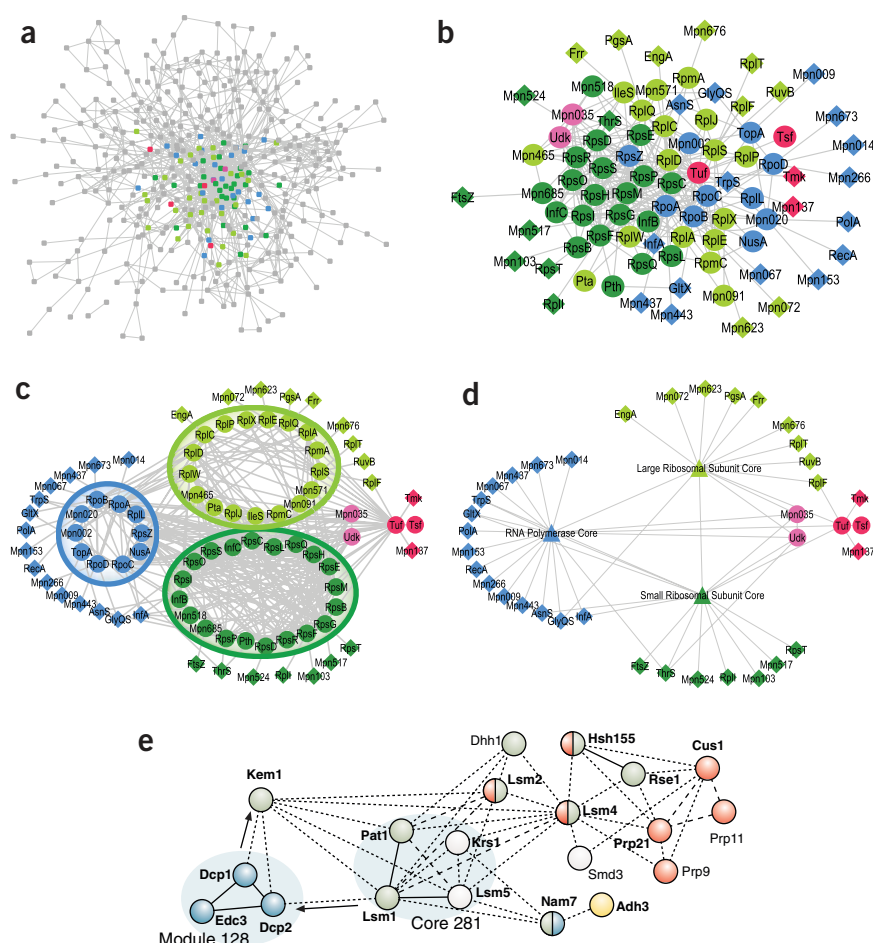
Name	Cost	OS	Description	URL
<b>Stand-alone</b>				
Arena 3D <sup>63</sup>	Free	Win, Mac, Linux	Visualization of biological multi-layer networks in 3D	<a href="http://www.arena3d.org/">http://www.arena3d.org/</a>
BiNA <sup>81</sup>	Free	Win, Mac, Linux	Exploration and interactive visualization of pathways	<a href="http://www.bnplusplus.org/bina/">http://www.bnplusplus.org/bina/</a>
BioLayout Express 3D <sup>37</sup>	Free	Win, Mac, Linux	Generation and cluster analysis of networks with 2D/3D visualization	<a href="http://www.biolayout.org/">http://www.biolayout.org/</a>
BiologicalNetworks <sup>82</sup>	Free	Win, Mac, Linux	Analysis suite; visualizes networks and heat map; abundance data	<a href="http://www.biologicalnetworks.org/">http://www.biologicalnetworks.org/</a>
Cytoscape <sup>*20,83</sup>	Free	Win, Mac, Linux	Network analysis; extensive list of plug-ins for advanced visualization	<a href="http://www.cytoscape.org/">http://www.cytoscape.org/</a>
GENeVis <sup>36</sup>	Free	Win, Mac, Linux	Network and pathway visualization; abundance data	<a href="http://tinyurl.com/genevis/">http://tinyurl.com/genevis/</a>
Medusa <sup>84</sup>	Free	Win, Mac, Linux	Basic network visualization tool	<a href="http://coot.embl.de/medusa/">http://coot.embl.de/medusa/</a>
N-Browse <sup>85</sup>	Free	Win, Mac, Linux	Network visualization software for heterogeneous interaction data	<a href="http://www.gnetbrowse.org/">http://www.gnetbrowse.org/</a>
NAVIGATOR <sup>23,86</sup>	Free	Win, Mac, Linux	Visualization of large protein-interaction data sets; abundance data	<a href="http://tinyurl.com/navigator1/">http://tinyurl.com/navigator1/</a>
Ondex <sup>87</sup>	Free	Win, Mac, Linux	Integrative workbench: large network visualizations; abundance data	<a href="http://www.ondex.org/">http://www.ondex.org/</a>
Osprey <sup>88</sup>	Free	Win, Mac, Linux	Tool for visualization of interaction networks	<a href="http://tinyurl.com/osprey1/">http://tinyurl.com/osprey1/</a>
Pajek <sup>89</sup>	Free	Win	Generic network visualization and analysis tool	<a href="http://pajek.imfm.si/">http://pajek.imfm.si/</a>
ProViz	Free	Win, Mac, Linux	Software for visualization and exploration of interaction networks	<a href="http://tinyurl.com/proviz/">http://tinyurl.com/proviz/</a>
SpectralNET <sup>90</sup>	Free	Win	Network visualizations; scatter plots for dimensionality reduction methods	<a href="http://tinyurl.com/spectralnet/">http://tinyurl.com/spectralnet/</a>
Tulip <sup>91</sup>	Free	Win, Mac, Linux	Generic visualization tool; extremely large networks; 3D support	<a href="http://tulip.labri.fr/TulipDrupal/">http://tulip.labri.fr/TulipDrupal/</a>
VANTED <sup>21</sup>	Free	Win, Mac, Linux	Combined visualization of abundance data, networks and pathways	<a href="http://tinyurl.com/vanted/">http://tinyurl.com/vanted/</a>
yEd	Free	Win, Mac, Linux	Generic network visualization software; offers many layout algorithms	<a href="http://tinyurl.com/yEdGraph/">http://tinyurl.com/yEdGraph/</a>
<b>Cytoscape plug-in</b>				
BiNoM <sup>92</sup>	Free	Win, Mac, Linux	Extensive support for common systems biology network formats	<a href="http://tinyurl.com/binom1/">http://tinyurl.com/binom1/</a>
BioModules <sup>24</sup>	Free	Win, Mac, Linux	Detects modules in networks; maps abundance data onto nodes and modules	<a href="http://tinyurl.com/biomodules/">http://tinyurl.com/biomodules/</a>
Cerebral <sup>*26,78</sup>	Free	Win, Mac, Linux	Biologically motivated layout algorithm; maps abundance data; clustering	<a href="http://tinyurl.com/cerebral1/">http://tinyurl.com/cerebral1/</a>
MCODE <sup>18</sup>	Free	Win, Mac, Linux	Network clustering algorithm; support for manual cluster refinement	<a href="http://tinyurl.com/MCODE123/">http://tinyurl.com/MCODE123/</a>
VistaClara <sup>42</sup>	Free	Win, Mac, Linux	Mapping of abundance data to nodes and 'heat strips'; provides heat map	<a href="http://tinyurl.com/cytoplugins/">http://tinyurl.com/cytoplugins/</a>
<b>Web-based</b>				
Graphle <sup>93</sup>	Free		Distributed client/server network exploration and visualization tool	<a href="http://tinyurl.com/graphle/">http://tinyurl.com/graphle/</a>
Lichen	Free		Library for web-based visualization of network and abundance matrix data	<a href="http://tinyurl.com/Lichen1/">http://tinyurl.com/Lichen1/</a>
MAGGIE Data Viewer	Free		Visualization of networks; abundance data in heat maps and profile plots	<a href="http://maggie.systemsbiology.net/">http://maggie.systemsbiology.net/</a>
STITCH <sup>31</sup>	Free		Construction and visualization of networks from a wide range of sources	<a href="http://stitch.embl.de/">http://stitch.embl.de/</a>
VisANT <sup>22</sup>	Free	Win, Mac, Linux	Analysis, mining and visualization of pathways and integrated omics data	<a href="http://visant.bu.edu/">http://visant.bu.edu/</a>

Some of the tools in this table have capabilities similar to tools that are listed in other tables. To avoid listing tools in more than one table, we assigned tools to tables on the basis of what we understand to be their primary purpose. \*Our recommendations. Free means the tool is free for academic use; \$ means there is a cost. OS, operating system: Win, Microsoft Windows; Mac, Macintosh OS X. Tools running on Linux usually also run on other versions of Unix. 2D, two-dimensional; 3D, three-dimensional.

to minimize the number of edge crossings while trying to keep the lengths of all edges approximately the same. However, this approach quickly becomes inadequate as the network size and complexity increase (Fig. 1a). Instead, clustering approaches are used, which predict higher-order protein complexes from the interaction data. One very commonly used tool for this purpose is MCODE<sup>18</sup>. For TAP-MS and other data sets where components of protein complexes are experimentally determined, other clustering methods are used (for example, 'clique percolation'<sup>19</sup>). The results of these clustering analyses can then be used to change the layout and appearance of the network (Fig. 1a,b) in a way that may yield biological insights that cannot be easily obtained by simply examining lists of proteins or protein complexes. For instance, by viewing the network, the scientist may notice connections between two complexes that suggest a previously unknown biological relationship. Furthermore, on the basis of previous knowledge, the scientist may be able to assign a putative function or subcellular localization to the complex; this information can be visualized using node color or shape to represent the functional category or location of the proteins. Similarly, node color or shape can be used to show which proteins belong to the same complex (Fig. 1b).

Most network visualization tools provide the ability to interactively change the layout of the network—for example, by automatically arranging a user-defined group of proteins into any of a variety of arrangements (a circle, a line and so forth) or by manually moving nodes. This ability can be very useful in creating visualizations that emphasize biologically significant relationships and interactions between complexes (Fig. 1c) or between 'hub' proteins and their partners (for example, between kinases and their substrates). Tools that support such interactive editing particularly well include Cytoscape<sup>20</sup>, VANTED<sup>21</sup>, VisANT<sup>22</sup> and NAViGaTOR<sup>23</sup>.

It is often useful to collapse all members of a protein complex or cluster into a single 'meta-node' (Fig. 1d) that can later be expanded, depending on screen space and the desired level of detail. Meta-nodes not only simplify the appearance of the network, they can also be useful in more clearly illustrating biological relationships between protein complexes. Meta-nodes can also help to visually arrange the network to give insight into the integration and coordination of cellular functions (Fig. 1d). Meta-nodes are supported by yEd (<http://tinyurl.com/yEdGraph/>), BioModules<sup>24</sup> and VisANT, the last of which further allows meta-nodes to be



**Figure 1** | Visualization of protein interaction networks. (a–d) Cytoscape<sup>20</sup> images of *Mycoplasma pneumoniae* protein interaction data derived by mass spectrometry<sup>19</sup> analysis. (a) Initial protein interaction network (>400 proteins) laid out with a force-directed algorithm. Nodes discussed in the following steps are overlaid with functional annotations (blue, RNA polymerase; dark or light green, small or large ribosomal subunits, respectively; red, elongation factor). (b) Recomputed network remaining after removal of nodes not of interest. Five computationally determined complexes are colored according to functional annotation. Node shapes represent different roles in the complex (circle, core protein of complex; diamond, protein attached to complex but not part of the core). At this stage, clusters emerge. (c) Manual refinement of the network layout emphasizing structure of protein complexes and interactions between them. (d) Collapse of nodes in each complex core, simplifying the network and emphasizing global properties. (e) Stages in deadenylation-dependent mRNA degradation in *Saccharomyces cerevisiae*. Reproduced from Gavin *et al.*<sup>10</sup>. Arrows show the order of sequential steps in a cellular process. Proteins are colored according to their localization (green, cytoplasm; red, nucleus; blue, punctate composite (undefined subcellular structure); yellow, mitochondria; white, unknown). Edge styles represent socio-affinity indices (dotted, 5–10; dashed, 10–15; solid, >15). TAP-MS bait proteins, bold; shaded circles, protein complexes.

nested hierarchically and can show 'meta-edges'<sup>25</sup> between meta-nodes—these can indicate, for example, when proteins are shared between two collapsed complexes.

Present high-throughput experimental methods often do not determine the spatial, or subcellular, location where an interaction takes place, so it can be highly informative to include any previous protein localization information in the analysis of these data sets. For instance, the network may be filtered to show only proteins known to occur in selected locations, thus simplifying it and allowing the scientist to focus only on interactions within a defined subcellular location. Alternatively, subcellular location



**Table 2** | Visualization tools focused on pathways

Name	Cost	OS	Description	URL
<b>Stand-alone</b>				
BioTapestry <sup>94</sup>	Free	Win, Mac, Linux	Visualization of genetic regulatory networks, also with experimental data	<a href="http://www.biotapestry.org/">http://www.biotapestry.org/</a>
Caleydo <sup>95</sup>	Free	Win, Linux	Interactive framework for pathway and expression data; 3D 'bucket' view	<a href="http://www.caleydo.org/">http://www.caleydo.org/</a>
CellDesigner <sup>*51</sup>	Free	Win, Mac, Linux	Drawing and simulation of pathways and models; supports SBGN	<a href="http://www.celldesigner.org/">http://www.celldesigner.org/</a>
Edinburgh Pathway Editor	Free	Win, Mac, Linux	Construction and visualization of pathway diagrams; supports SBGN	<a href="http://tinyurl.com/EdinburghPE/">http://tinyurl.com/EdinburghPE/</a>
GenMAPP <sup>40</sup>	Free	Win	Pathway visualization and construction; abundance data	<a href="http://www.genmapp.org/">http://www.genmapp.org/</a>
IngenuityPathways	\$	Win, Mac, Linux	Full analysis suite; network and pathway visualizations; abundance data	<a href="http://tinyurl.com/IngenuityPath/">http://tinyurl.com/IngenuityPath/</a>
JDesigner <sup>52</sup>	Free	Win	Drawing and simulation of pathways and models	<a href="http://tinyurl.com/jdesigner/">http://tinyurl.com/jdesigner/</a>
KaPPA View <sup>48</sup>	Free	Win	Analysis and visualization of plant pathways and mapped abundance data	<a href="http://tinyurl.com/kappa-view/">http://tinyurl.com/kappa-view/</a>
KEGG Atlas <sup>96</sup>	Free	Win, Mac, Linux	Visualization of abundance data on interactive KEGG pathways	<a href="http://www.genome.jp/kegg/">http://www.genome.jp/kegg/</a>
MetaCore	\$	Win, Mac, Linux	Pathway, network and omics data analysis and visualization suite	<a href="http://www.genego.com/">http://www.genego.com/</a>
PathVisio <sup>97</sup>	Free	Win, Mac, Linux	Pathway visualization and editing; supports mapping of omics data	<a href="http://www.pathvisio.org/">http://www.pathvisio.org/</a>
VitaPad <sup>98</sup>	Free	Win, Mac, Linux	Editing of pathway diagrams; integration of abundance data	<a href="http://tinyurl.com/vitapad/">http://tinyurl.com/vitapad/</a>
<b>Web-based</b>				
ArrayXPath <sup>99</sup>	Free		Mapping of abundance data to pathway visualizations	<a href="http://tinyurl.com/ArrayXPath/">http://tinyurl.com/ArrayXPath/</a>
GEPAT <sup>100</sup>	Free		Analysis suite; visualization of transcriptomics data on pathways maps	<a href="http://tinyurl.com/GEPAT1/">http://tinyurl.com/GEPAT1/</a>
iPath <sup>101</sup>	Free		Visualization and exploration of combined KEGG pathways	<a href="http://pathways.embl.de/">http://pathways.embl.de/</a>
MapMan <sup>46</sup>	Free		Visualization of abundance data on metabolic pathways	<a href="http://tinyurl.com/MapManApp/">http://tinyurl.com/MapManApp/</a>
Omics Viewer <sup>47,102</sup>	Free		Mapping of abundance data to BioCyc pathway diagrams	<a href="http://www.biocyc.org/">http://www.biocyc.org/</a>
Pathway Explorer <sup>49</sup>	Free		Visualization of abundance data on pathways	<a href="http://tinyurl.com/pathwayexp/">http://tinyurl.com/pathwayexp/</a>
PATIKA <sup>103</sup>	Free		Pathway visualization suite; good support for signaling pathways	<a href="http://www.patika.org/">http://www.patika.org/</a>
Payaologue	Free		Collaborative pathway annotation and visualization tool	<a href="http://celldesigner.org/payao/">http://celldesigner.org/payao/</a>
ProMeTra <sup>41</sup>	Free		Maps abundance matrices of multiple data types to pathways	<a href="http://tinyurl.com/ProMeTra/">http://tinyurl.com/ProMeTra/</a>
Reactome SkyPainter <sup>30</sup>	Free		Visualization of over-represented pathways and reactions from gene lists	<a href="http://reactome.org/">http://reactome.org/</a>
WikiPathways <sup>62</sup>	Free		Wiki-based, community-driven pathway curation and visualization tool	<a href="http://www.wikipathways.org/">http://www.wikipathways.org/</a>

Some of the tools in this table have capabilities similar to tools that are listed in other tables. To avoid listing tools in more than one table, we assigned tools to tables on the basis of what we understand to be their primary purpose. \*Our recommendations. Free means the tool is free for academic use; \$ means there is a cost. OS, operating system: Win, Microsoft Windows; Mac, Macintosh OS X. Tools running on Linux usually also run on other versions of Unix. 3D, three-dimensional.

can be indicated using node coloring; this can be particularly useful when studying the interactions of complexes that move between subcellular locations (Fig. 1e). Another common strategy is to arrange the network so that all proteins belonging to the same subcellular location are gathered together in one region (Fig. 2, see 2a). For small networks, such a layout depicting subcellular localization is often created manually. However, for large networks, it is much more convenient to use tools that can achieve such a layout automatically, such as Cerebral<sup>26</sup> (Fig. 2a) and PATIKA<sup>27</sup>. These tools also draw boundaries or use shading so that the scientist can see clearly which regions of the network correspond to which subcellular locations.

Protein interaction data sets commonly do not capture information about dynamic changes in protein abundance. Thus, as with spatial information, it is often useful to include temporal information from other experiments—for example, by identifying proteins whose abundance is known to vary throughout the cell cycle. This information can be used to simplify a large network by either depicting only proteins that are coexpressed or by mapping expression or abundance profiles of proteins of interest onto nodes<sup>28</sup> in the network, as described in more detail below (see Network enrichment).

These processes of dissection are all aimed at dividing a protein interaction data set into manageable, biologically significant parts that can be interpreted; during this process of interpretation, a scientist often makes use of previously established knowledge, particularly pathways (for example, KEGG<sup>29</sup> or Reactome<sup>30</sup>) and

networks (for example, STITCH<sup>31</sup>). In some cases, to illustrate a result or insight, it can be useful to add interactions derived from previous studies—thus forming a hybrid network that shows both new and old data (Fig. 1e).

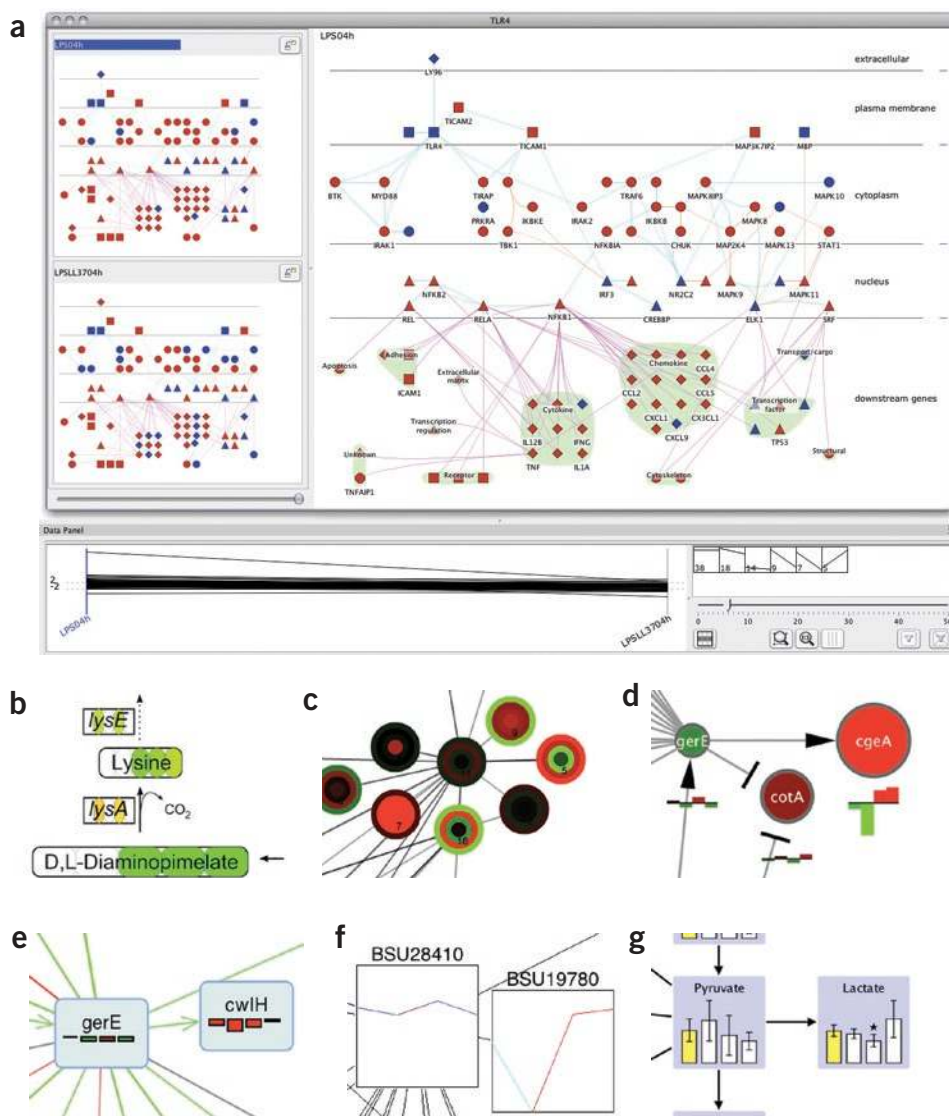
### Expression profile data

A range of experimental methods are being used for high-throughput expression profiling (Box 1, Supplementary Fig. 1 and Supplementary Table 1); in addition to gene expression profiling with DNA microarrays<sup>32</sup> and RNA deep sequencing<sup>33</sup>, a promising emerging technology is quantitative protein expression profiling based on mass spectrometry<sup>34,35</sup>. Gene expression profile data sets are being deposited in two main repositories, ArrayExpress (<http://www.ebi.ac.uk/arrayexpress/>) and Gene Expression Omnibus (<http://www.ncbi.nlm.nih.gov/geo/>), with around 15,000 studies now in the public domain.

The initial goal in analyzing expression profiles is usually to find a set of genes or, less typically, proteins that share a related pattern of expression—for example, genes that are up- or down-regulated in a certain genotype, disease model or human disease, or in response to a drug treatment. The challenge is that a single data set may contain expression profiles for over 10,000 genes, measured over a range of time points and experimental conditions, so that determining which genes are potentially relevant to the studied problem requires an extensive search through a large amount of often noisy, multivariate data. Together with various clustering algorithms, visualization is key in these analyses<sup>32</sup>, and

**Figure 2** | Omics data overlaid onto biological networks. (a) Cerebral<sup>78</sup> showing the TLR4-to-NF- $\kappa$ B signaling pathway<sup>79</sup> laid out according to subcellular localization and functional annotation (green shading). Direction of information flow is from top to bottom. Node colors represent relative expression (red, upregulation; blue, downregulation) and edge colors represent interaction type (orange, phosphorylation; cyan, other protein interaction; purple, transcriptional regulation). The left two panels are a 'small multiples' display of the same pathway overlaid with gene expression data for two experimental conditions. In this case, the top panel has been selected, and hence is also shown in the main window. The bottom panels show the detailed expression profiles corresponding to genes shown in the pathway panel (see 'Network enrichment'). The data set shows how upregulation of NFKB1 explains the observed upregulation of several chemokine proteins.

(b) ProMeTra<sup>41</sup> display showing both metabolomics and transcriptomics time series data from five time points. Metabolite and enzyme nodes in the pathway map are subdivided into five areas, one per time point. Areas are color coded (green, upregulation; yellow, no change; white, missing data) to indicate metabolite concentrations and transcript levels relative to a reference time point. (c) Lichen rendering of a gene regulatory network overlaid with transcriptomics data using a circular heat map. Each concentric ring represents a time point, and the color of the circle represents expression level (red, upregulation; green, downregulation; black, no change). Numbers identify genes. (d) VistaClara<sup>42</sup> display showing transcript levels across four time points relative to a reference time point as 'heat strips' below the nodes (height of bar, relative expression; red, upregulation; green, downregulation). Node color indicates expression level from time point 4. Node size indicates reliability of the measurement taken at time point 4. (e) GENEVis<sup>36</sup> visualization of the same data set. Color coding as in d; height of bar corresponds directly to the reliability of the measurement at each time point. (f) Profile plots of the same data embedded in nodes in VisANT<sup>22</sup>. The color of each line segment represents the change in expression levels between two time points (red and purple, increase; blue and cyan, decrease). (g) Visualization of metabolite concentrations in a pathway map in VANTED<sup>21</sup> using a bar chart with error bars.



a wide range of tools have been developed to aid the visualization process (Table 3). Many of these tools implement a set of commonly applied methods (Box 2 and Fig. 3); in particular, scatter plots combined with dimensionality reduction (Fig. 3a), profile plots (Fig. 3b), heat maps, and dendrograms (Fig. 3c), as well as clustering. As microarray gene expression analysis has matured as an experimental technique, many of the corresponding visualization methods have become well established and are widely used.

**Network enrichment.** Once a list of potentially relevant genes has been found using the above types of analysis, the next task is often to find pathways or networks where these genes are significantly over-represented. These 'enrichment' searches can be launched directly from several network visualization tools, for example, GENEVis<sup>36</sup>, Reactome SkyPainter<sup>30</sup>, Metacore (GeneGo Inc.)

or BioLayout Express 3D<sup>37</sup>. A logical next step is then to map gene expression levels onto the identified pathways. Interpreting expression data in the context of a visualized pathway or network usually proves more insightful than without this type of information. For instance, visualizing the data in the context of pathways may show how the upregulation of a transcription factor explains the upregulation of many other genes under its control (Fig. 2a) and may lead to testable experimental hypotheses.

A wide range of representations are used for mapping gene expression levels onto pathways and networks, with the ideal choice depending on the specific experiment and question of interest<sup>38</sup> (Fig. 2). A simple approach that is available as part of many tools (Table 1) is to represent expression levels as a color gradient, as in a heat map, and then color the nodes in the network according to their expression level under a particular condition (Fig. 2a).

**Table 3** | Visualization tools for multivariate omics data

Name	Cost	OS	Description	URL
<b>Stand-alone</b>				
BicOverlapper <sup>104</sup>	Free	Win, Mac, Linux	Visualization of biclusters combined with profile plots and heat maps	<a href="http://vis.usal.es/bicoverlapper/">http://vis.usal.es/bicoverlapper/</a>
BiGGEsTS <sup>105</sup>	Free	Win, Mac, Linux	Heat map-based bicluster visualization	<a href="http://tinyurl.com/BiGGEsTS/">http://tinyurl.com/BiGGEsTS/</a>
Brain Explorer <sup>76</sup>	Free	Win, Mac	Visualization of 3D transcription data in the central nervous system	<a href="http://tinyurl.com/brainExplorer/">http://tinyurl.com/brainExplorer/</a>
Caryoscope <sup>75</sup>	Free	Win, Mac, Linux	Abundance data mapped to chromosomal location	<a href="http://tinyurl.com/caryoscope/">http://tinyurl.com/caryoscope/</a>
Data Matrix Viewer	Free	Win, Mac, Linux	Simple profile plot visualization; supports Gaggles	<a href="http://gaggles.systemsbio.net/">http://gaggles.systemsbio.net/</a>
EXPANDER <sup>106</sup>	Free	Win, Linux	Heat maps, scatter plots and profile plots of cluster averages	<a href="http://acgt.cs.tau.ac.il/expander/">http://acgt.cs.tau.ac.il/expander/</a>
Genesis <sup>107</sup>	Free	Win, Mac, Linux	Analysis suite; offers several interactive visualizations	<a href="http://tinyurl.com/genesclient/">http://tinyurl.com/genesclient/</a>
GeneSpring GX*	\$	Win, Mac, Linux	Analysis suite; interactive and linked visualizations; also networks	<a href="http://tinyurl.com/genespring/">http://tinyurl.com/genespring/</a>
GeneVAnD <sup>108</sup>	Free	Win, Mac, Linux	Linked heat maps, dendrograms and 2D/3D scatter plots	<a href="http://tinyurl.com/GeneVAnD/">http://tinyurl.com/GeneVAnD/</a>
geWorkbench	Free	Win, Mac, Linux	Modular suite; heat maps, dendrograms, profile and scatter plots	<a href="http://tinyurl.com/geWorkbench/">http://tinyurl.com/geWorkbench/</a>
HCE* <sup>109</sup>	Free	Win	Linked heat map, profile and scatter plots; systematic exploration	<a href="http://tinyurl.com/HCEplorer/">http://tinyurl.com/HCEplorer/</a>
Java TreeView* <sup>110</sup>	Free	Win, Mac, Linux	Linked heat maps, karyoscopes, sequence alignments, scatter plots	<a href="http://jtreeview.sourceforge.net/">http://jtreeview.sourceforge.net/</a>
Mayday <sup>111</sup>	Free	Win, Mac, Linux	Modular suite; many linked visualizations; enhanced heat map <sup>112</sup>	<a href="http://tinyurl.com/maydaywp/">http://tinyurl.com/maydaywp/</a>
MultiExperiment Viewer* <sup>113</sup>	Free	Win, Mac, Linux	Analysis suite; heat maps, dendrograms, profile and scatter plots	<a href="http://www.tm4.org/">http://www.tm4.org/</a>
PointCloudXplore <sup>77</sup>	Free	Win, Mac, Linux	Visualization of 3D transcription data in <i>Drosophila</i> embryos	<a href="http://tinyurl.com/PointCloudXplore/">http://tinyurl.com/PointCloudXplore/</a>
Spotfire Functional Genomics	\$	Win	Analysis suite; many linked visualizations and exploration tools	<a href="http://spotfire.tibco.com/">http://spotfire.tibco.com/</a>
TimeSearcher <sup>114</sup>	Free	Win	Exploration and analysis of time series; advanced profile plots	<a href="http://tinyurl.com/timesearcher/">http://tinyurl.com/timesearcher/</a>
<b>R/BioConductor</b>				
Geneplotter	Free	Win, Mac, Linux	Karyoscope-style plots and other visualizations	<a href="http://www.bioconductor.org/">http://www.bioconductor.org/</a>
<b>Web-based</b>				
ExpressionProfiler <sup>115</sup>	Free		Transcriptomics data analysis suite with basic visualizations	<a href="http://tinyurl.com/exprespro/">http://tinyurl.com/exprespro/</a>
GenePattern <sup>116</sup>	Free		Modular analysis platform; several visualization modules available	<a href="http://tinyurl.com/GenePatt/">http://tinyurl.com/GenePatt/</a>

Some of the tools in this table have capabilities similar to tools that are listed in other tables. To avoid listing tools in more than one table, we assigned tools to tables on the basis of what we understand to be their primary purpose. \*Our recommendations. Free means the tool is free for academic use; \$ means there is a cost. OS, operating system: Win, Microsoft Windows; Mac, Macintosh OS X. Tools running on Linux usually also run on other versions of Unix. 2D, two-dimensional; 3D, three-dimensional.

If expression levels from more than one condition are being studied, some tools (for example, VisANT and VistaClara) allow the scientist to visualize them sequentially, by updating node colors to reflect the expression levels of a selected condition. Some tools switch automatically to depiction of the next condition after a predefined time interval, which leads to an animation-like visualization that is well suited to interpreting data from a time series. An alternative strategy to viewing the data from different conditions in series is to view them in parallel, by arranging multiple versions of the same network in a grid, where each version represents the expression levels (visualized as node color) for one condition or time point. This approach is known as ‘small multiples’<sup>39</sup> and allows the scientist to visually compare expression levels between conditions, which is not well supported by animation. A well-designed implementation of small multiples is available in Cerebral<sup>26</sup> (Fig. 2a).

Besides animation and small multiples, a third approach is to show the complete expression profile within the nodes of a network. The most common representation of this type is based on a miniature heat map embedded in each node (Fig. 2b) and is available in several tools, including GenMAPP 2 (ref. 40), GeneSpring GX (Agilent Technologies) and ProMeTra<sup>41</sup>. The Lichen package (<http://tinyurl.com/Lichen1/>) uses a circular heat map to depict this information, which has the advantage of being very compact (Fig. 2c). VistaClara<sup>42</sup> provides ‘heat strips’, in which the heights of the bars as well as their colors correspond to the expression levels (Fig. 2d). In contrast, in GENeVis, bar heights represent confidence measures, so that reliable measurements are taller

and therefore emphasized (Fig. 2e). A less common alternative to showing a heat map embedded in the node is to embed a profile plot in the node. This is supported by, for instance, VisANT (Fig. 2f) and has the advantage that multiple profiles can be displayed in the same node—for example, when a meta-node represents a set of genes. In VANTED, each node of the network has embedded visualizations of exceptionally high detail, showing legends, grid lines, bar charts or error bars (Fig. 2g). Although powerful, this representation requires the nodes to be rather large in order to show the details of the embedded visualizations, which effectively limits its application to only small pathways and networks.

For expression profiles with many conditions, visualizing all these data directly in the network is invariably problematic because of a lack of space, and an approach that links visualization of the network to a separate visualization of the expression profiles is more appropriate. In the linked approach, a heat map (as implemented in VistaClara) or a profile plot (as implemented in Cerebral; Fig. 2a) is shown next to the network and when the scientist selects nodes in the network, the corresponding expression profiles are highlighted in the linked heat map or profile plot, or vice versa. This approach allows the scientist to check, for instance, whether the members of a putative protein complex in the network visualization are coexpressed, by comparing the corresponding gene expression profiles in the linked heat map. Conversely, selection of a set of coexpressed genes in a clustered heat map would allow exploration of their role in the linked protein interaction network: the scientist could directly see whether these genes are part of the same complex, what their interaction

## BOX 2 KEY VISUALIZATION METHODS FOR MULTIVARIATE DATA

Multivariate data, for instance from gene expression studies, are very common in systems biology, and many tools have been developed to analyze and visualize such data (Table 3). The three most commonly used visualization methods are scatter plots, profile plots and heat maps.

**Scatter plots.** Scatter plots (Fig. 3a) are primarily used to examine dependencies between two variables, but in combination with dimensionality reduction methods, they can also be applied to multivariate data. For instance, to gain insight into the global patterns in a gene expression matrix, a dimensionality reduction method may be applied to obtain a two-dimensional (sometimes three-dimensional) representation of the expression profiles, which are then visualized in a scatter plot to reveal clusters and outliers in the data. Some frequently applied dimensionality reduction methods for this purpose are principal component analysis<sup>132</sup> (PCA) and multi-dimensional scaling<sup>133</sup> (MDS), which are implemented in many tools. Besides PCA and MDS, many other suitable dimensionality reduction methods exist<sup>134</sup>, but they are often not easily accessible to the casual user.

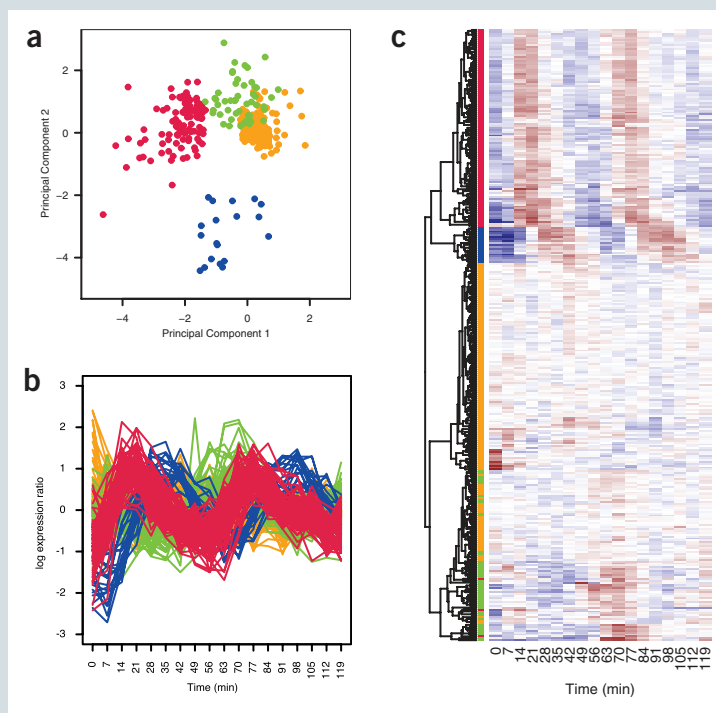
Scatter plots combined with dimensionality reduction methods are an excellent tool for gaining insight into the overall structure of large sets of expression profiles. However, because of the dimensionality reduction itself, it is not possible to extract information about the relationship between expression levels and the conditions under study.

**Profile plots.** Profile plots (Fig. 3b), also known as parallel coordinate plots<sup>135</sup>, visualize the expression levels of a large number of transcripts across all samples. Thus, they provide insight into the patterns of correlation between samples and expression levels. For instance, at a glance, the scientist can determine whether a transcript is expressed constitutively in all conditions or whether it is only expressed in a single condition, such as a particular tissue or phase of the cell cycle. Furthermore, it is possible to generate hypotheses about trends, such as increasing expression levels for a transcript over time after a stimulus, or differential expression of a transcript—for instance, between samples of diseased and normal tissue. Because many profiles are shown in the same plot, the scientist can interpret such observations in the context of the overall data set.

A profile plot can also be queried visually for transcripts with a particular behavior, such as low expression in one set of samples and high in another set, or for profiles that are similar to that of a transcript of interest. A substantial disadvantage of profile plots is that, owing to the manner in which they are constructed, profiles overlap, severely limiting the number of profiles that can be visualized effectively at the same time.

**Heat maps.** Heat maps<sup>136,137</sup> (Fig. 3c) are the most commonly used visualization method for expression matrices<sup>138</sup> and can be generated using most tools. Like profile plots, heat maps visualize the abundance of each transcript in each sample, but the profiles do not overlap, which means that more profiles can be visualized effectively. However, the size of the heat map grows with the number of profiles, so that the available screen space is often a limiting factor.

A key aspect of heat map visualization is the reordering of the rows, which ensures that similar profiles are placed near each other. Typically this reordering is done using hierarchical clustering<sup>137</sup>, and a dendrogram showing the hierarchy is usually arranged immediately adjacent to the heat map (Fig. 3). This combined view helps a scientist to see groups of genes that have a similar expression pattern. The dendrogram conveys which genes are clustered together, and also which genes are outliers with an unusual expression pattern. The heat map allows the scientist to see in more detail which features of the expression pattern are shared by gene clusters. For example, genes in a cluster may have a peak expression at about the same time in an experiment.



**Figure 3** | Visualization of gene expression profiles. Expression of 320 transcripts from *S. cerevisiae*, collected over 18 time points throughout the cell cycle<sup>80</sup>. Colors indicate cluster membership based on a *k*-means clustering (*k* = 4). (a) Scatter plot showing a projection of the profiles on the first two principal components obtained by PCA. (b) Profile plot of gene expression across all 18 time points, including *k*-means cluster information. Genes in the red and blue clusters appear active in the G1 and S phase of the cell cycle, respectively. Phase assignments for yellow and green clusters are unclear. (c) Heat map of the profiles. Colors represent abundance (red, higher than control; blue, lower than control; white, no change). Rows of the heat map have been reordered according to a hierarchical clustering, represented by the dendrogram. The color bars between the dendrogram and heat map indicate the *k*-means clusters, allowing comparison of the two clustering results. Images made with R (<http://www.r-project.org/>).



partners are, or whether they are located in the same subcellular compartment. In contrast, when expression profiles are shown only in the nodes of the network, this type of analysis is not possible because coexpressed genes are not necessarily located next to each other in the visualization. However, there is a trade-off between the flexibility provided by linked views and the convenience of being able to see expression profiles and interactions without having to consult two separate visualizations.

**Network clustering and correlation networks.** Recently, there has been increased interest in a new kind of clustering method—called ‘network clustering’<sup>37</sup>—that is less susceptible to noise and can lead to more accurate identification of functionally related genes than established clustering methods (Box 2). Network clustering of gene expression data is done using so-called ‘correlation networks’, in which each gene is a node and each edge indicates coexpression of two genes under the conditions of the experiment<sup>43</sup>. As well as being an improved way to calculate clusters, correlation networks allow the scientist to interactively explore gene expression data sets using many of the rich set of network visualization tools that have been developed for visualizing protein interaction networks. The use of correlation networks for gene expression data is as yet supported by relatively few tools, including BioLayout Express 3D<sup>37</sup>—which has been developed specifically for this purpose—and Cytoscape, using either the MCODE<sup>18</sup> or ClusterMaker plug-in (<http://www.cgl.ucsf.edu/cytoscape/>). However, we anticipate that correlation networks may become one of the established methods for interpreting gene expression data sets.

### Metabolic profile data

A wide variety of spectroscopic methods are being used for high-throughput studies of small-molecule metabolites<sup>44</sup>, two of the most popular being mass spectrometry and nuclear magnetic resonance spectroscopy (Box 1 and Supplementary Table 1). Typically, present methods identify hundreds of metabolites per experiment. Additionally, many as-yet-unidentified compounds can be reproducibly detected. These experimental data are collected in several public repositories, the largest of which is now SetupX<sup>45</sup>, containing ~20,000 samples from more than 300 studies.

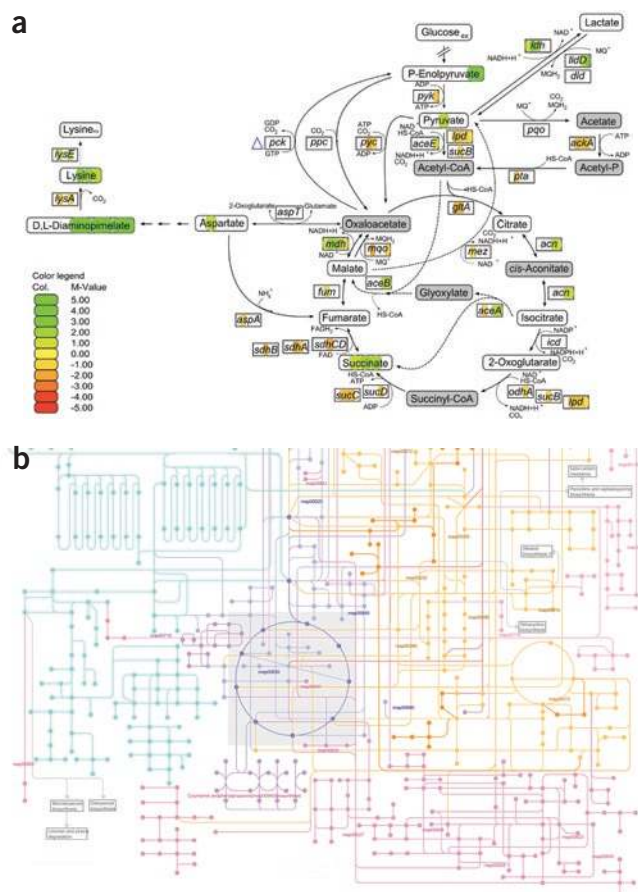
The general goal in analyzing metabolite profiles is to gain detailed insight into the molecular mechanisms of cellular metabolic pathways. The identification of molecules that may be used as reliable biomarkers of disease is also of great interest. Metabolite profiles are typically analyzed to find sets of metabolites with similar profiles or to measure the impact of genetic modifications, drugs and other biotic or abiotic factors on the metabolome of an organism.

As with gene and protein profiles, visualization is key in these analyses, and the same, or very similar, methods (Box 2) and tools (Table 3) are typically used. As for gene expression data, one of the key visualization methods in metabolomics involves the enrichment of metabolic pathways with visualizations of metabolite concentrations (Fig. 4a), and often the same visual representations as for gene expression data can be used (Fig. 2b–g). Visualizing such enriched metabolic pathways can be very useful in understanding the concerted changes of metabolite pools within the cell. In addition, enriched pathways can help to identify metabolites that should be present according to enzymatic

reactions contained in the metabolic pathway but that have not been detected in the measurements. If such metabolites are identified, further experiments attempting to detect these molecules can be conducted. Many visualization tools (for example, MapMan<sup>46</sup>, Pathway Tools Omics Viewer<sup>47</sup>, KaPPA-view<sup>48</sup>, PathwayExplorer<sup>49</sup> and ProMeTra<sup>41</sup>) have been developed to facilitate enriched views of metabolic pathways, usually with close integration of metabolic pathway databases. These tools overlay metabolite profiles primarily on static images of pathways obtained from sources such as KEGG<sup>29</sup> or MetaCyc<sup>50</sup>-based databases.

### Pathway editing

The analysis of new experimental data sets, as outlined above, usually produces new insight into biological processes, which may be used to modify existing pathways or to create new pathways. A wide range of tools is available that support pathway editing (Table 2); the choice of which tool to use depends on the specific requirements of the task at hand.



**Figure 4** | Visualization of metabolic pathways and profile data. (a) A part of the glycolysis and citric acid cycle pathway in *Corynebacterium glutamicum* DM1730 overlaid with changes in metabolite concentrations and gene expression across five time points in relation to a reference time point. The visualization was created in ProMeTra<sup>41</sup> (as in Fig. 2b). Nodes shaded gray indicate metabolites for which no concentration data were available. (b) Enlarged iPath/KEGG Atlas image showing the glycolysis pathway in the context of other parts of the metabolic system. Yellow, amino acid metabolism, purple, energy metabolism. The shaded area corresponds to the citric acid cycle shown in a.

For building pathways from scratch or editing existing pathways, tools such as GenMAPP 2, PathVisio, and VANTED are useful, as they are designed for to assist the manual task of arranging nodes and edges. To this category also belong Cell Designer<sup>51</sup> and JDesigner<sup>52</sup>, which further support pathway simulations by means of kinetic modeling. The insights gained from these simulations often lead to new hypotheses, which can then be tested in further experiments.

Manual layout of pathways quickly becomes tedious as the size of the pathway grows. Fortunately, a range of automated layout methods have been developed, each addressing specific layout requirements. Typically, these methods will arrange the network to highlight the overall state changes that occur—for example, making sure that all interactions point from left to right, and thus creating an overall causal flow from left to right. Automated layout can be particularly useful for updating large networks when new knowledge (nodes or interaction edges) becomes available. For example, PATIKA<sup>53</sup> has an automated layout method that shows the causal flow of events through different subcellular compartments. This is particularly useful for depicting signaling networks<sup>27</sup>. Although these specialized automated layout methods are useful, they are usually of low quality compared to manually laid out pathways created by human experts and often require manual editing in addition; however, judging by recent progress, we expect these methods to continue to improve and to become increasingly useful<sup>54</sup>.

For very large pathways, it can be important to use compact visual representations and pathway layouts that reduce the amount of detail shown. A very clear illustration of such a concise visual representation is iPath, which combines 120 KEGG pathways into a single, vast pathway map that provides an overview of all metabolism in an organism (Fig. 4b). Scientists can zoom into parts of the map to navigate to individual pathways.

### Future perspectives

Systems biology is still rapidly evolving, which can make it difficult for tool developers to know which visualization tasks are the most important ones. However, as the field matures, the key tasks will likely become clearer, and the requirements and limitations of current visualization methods will become better understood<sup>5</sup>. This process will also be aided by insights from the emerging field of visual analytics<sup>55</sup>, which specifically studies the role of visualization in the larger process of understanding and interpreting data. Visual analytics methods have begun to be applied to studying the connection between visualization and analytical reasoning in systems biology<sup>5,56</sup>.

We anticipate that the near future will bring significant improvements in automated pathway and network layout to better match biologists' needs<sup>54,57</sup>. Innovation will continue to give more and better choices for the representations of nodes, edges and overlay information, as well as better ways to convey dynamic properties and to compare networks. Crucially, we expect that usability will improve, partly through improved navigation methods that help users manage large and complex networks<sup>23,25,58</sup>.

Today, many tools for network and pathway visualization are stand-alone applications (Tables 1 and 2); however, there is a trend toward web-based applications, often coupled tightly to underlying databases. Web-based tools show great promise for

facilitating collaboration between scientists at different locations<sup>59–61</sup>, and several projects have recently been launched that are aimed at community-based collaborative editing of biological network data—notably Payaologue (<http://celldesigner.org/payao/>) and WikiPathways<sup>62</sup>.

As experimental methods enable scientists to tackle larger and more complex systems, it is likely that significant innovations will be needed for visualizing future data sets. One possible direction for future network visualization tools would be to move beyond the standard two-dimensional layout, and tool developers are already exploring three-dimensional layouts (for example, BioLayout Express 3D<sup>37</sup>), combinations of both three-dimensional layouts and time (for example, E-Cell 3D, <http://tinyurl.com/ecell3d/>), or layouts that mix aspects of two and three dimensions (for example, Arena3D<sup>63</sup>). In addition, systems biologists may well be among the early adopters of innovations in hardware, such as multi-touch interfaces and larger, high-resolution displays<sup>64</sup>.

As systems biology has evolved very quickly over the last decade, some of the difficulties faced by end-users today arise not from the intrinsic complexity of data but from a lack of standards. Biological pathways and networks are now distributed in over 300 web resources<sup>4</sup>—and in a field as interdisciplinary as systems biology, there is an obvious strength in such diversity. However, the field would clearly benefit from a parallel effort toward a consolidated resource, and we would like to add our voices to a call for a consolidated database, similar to the worldwide Protein Data Bank for three-dimensional structures<sup>65</sup>.

The situation is somewhat better with file formats used to store interaction data, pathways and biochemical models. Although many formats are used, several have emerged as *de facto* standards for the exchange of pathway and network data—for example, PSI-MI<sup>66</sup> for protein interaction data, BioPAX (<http://www.biopax.org/>) for pathways and interaction networks, Systems Biology Markup Language (SBML)<sup>67</sup> for models of biochemical reactions and gene regulation and CellML<sup>68</sup> for exchange of a range of different biological models. In regard to graphical notation, there has recently been a significant community-driven proposal (Systems Biology Graphical Notation, SBGN<sup>69</sup>) toward developing a more unified standard, and several tools already support the creation and visualization of networks using this standard (see Table 2).

Ultimately, systems biology seeks to provide insights into the processes of organelles, cells, organs and even whole organisms. Fulfilling this ambitious goal requires still further development in visualization methods; in particular, better integration with visualization of other kinds of data, such as imaging data<sup>70</sup>, macromolecular structures<sup>71</sup>, genomes<sup>72</sup>, and phylogenies<sup>73</sup>. Efforts to build such integrated visualization platforms have begun (for example, Visible Cell<sup>74</sup>), and in fact, many tools that bridge different data types and disciplines are already in place; for instance, there are tools that map transcript abundance (or, if available, protein abundance) onto chromosomal location<sup>75</sup> and onto three-dimensional anatomical representations of tissue<sup>76,77</sup>. However, truly integrated visualization of systems biology data across the entire range of possible data types is still very much in its infancy.

*Note: Supplementary information is available on the Nature Methods website.*

# ACKNOWLEDGMENTS

The authors would like to acknowledge S. Kühner for providing the data for Figure 1 and Â. Gonçalves for comments on parts of the manuscript. This work was partly supported by the European Union Framework Programme 6 grant 'TAMAHUD' (LSHC-CT-2007-037472).

# COMPETING INTERESTS STATEMENT

The authors declare no competing financial interests.

Published online at <http://www.nature.com/naturemethods/>.

Reprints and permissions information is available online at <http://npg.nature.com/reprintsandpermissions/>.

1. Michal, G. *Biochemical Pathways: An Atlas of Biochemistry and Molecular Biology* (Wiley, New York, 1998).
2. Nishizuka, Y. The role of protein kinase C in cell surface signal transduction and tumour promotion. *Nature* **308**, 693–698 (1984).
3. Levine, M. & Davidson, E.H. Gene regulatory networks for development. *Proc. Natl. Acad. Sci. USA* **102**, 4936–4942 (2005).
4. Bader, G.D., Cary, M.P. & Sander, C. Pathguide: a pathway resource list. *Nucleic Acids Res.* **34** Database issue, D504–D506 (2006).
5. Saraiya, P., North, C. & Duca, K. Visualizing biological pathways: requirements analysis, systems evaluation, and research agenda. *Inf. Vis.* **4**, 191–205 (2005).

**This paper represents one of the first attempts to critically evaluate the requirements for visualization software used in biology.**

6. Suderman, M. & Hallett, M. Tools for visually exploring biological networks. *Bioinformatics* **23**, 2651–2659 (2007).
7. Pavlopoulos, G.A.G., Wegener, A.L.A. & Schneider, R.R. A survey of visualization tools for biological network analysis. *BioData Min.* **1**, 12 (2008).
8. Charbonnier, S., Gallego, O. & Gavin, A.C. The social network of a cell: recent advances in interactome mapping. *Biotechnol. Annu. Rev.* **14**, 1–28 (2008).
9. Yu, H. *et al.* High-quality binary protein interaction map of the yeast interactome network. *Science* **322**, 104–110 (2008).
10. Gavin, A.C. *et al.* Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**, 631–636 (2006).
11. Mathivanan, S. *et al.* An evaluation of human protein-protein interaction data in the public domain. *BMC Bioinformatics* **7** (suppl. 5), S19 (2006).
12. Ma'ayan, A. Network integration and graph analysis in mammalian molecular systems biology. *IET Syst. Biol.* **2**, 206–221 (2008).
13. Salwinski, L. *et al.* The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.* **32** (database issue), D449–D451 (2004).
14. Prasad, T.S., Kandasamy, K. & Pandey, A. Human Protein Reference Database and Human Proteinpedia as discovery tools for systems biology. *Methods Mol. Biol.* **577**, 67–79 (2009).
15. Aranda, B. *et al.* The IntAct molecular interaction database in 2010. *Nucleic Acids Res.* **38** (database issue), D525–D531 (2010).
16. von Mering, C. *et al.* Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**, 399–403 (2002).
17. Fruchterman, T.M.J. & Reingold, E.M. Graph drawing by force-directed placement. *Software Pract. Exper.* **21**, 1129–1164 (1991).
18. Bader, G.D. & Hogue, C.W. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* **4**, 2 (2003).
19. Kühner, S. *et al.* Proteome organization in a genome-reduced bacterium. *Science* **326**, 1235–1240 (2009).
20. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).

**This paper describes the core Cytoscape software, which has become one of the most popular tools to visualize and analyze biological networks. This is partly due to the modular design of the software that allows developers to create plug-ins to address virtually any network analysis problem.**

21. Junker, B.H., Klukas, C. & Schreiber, F. VANTED: a system for advanced data analysis and visualization in the context of biological networks. *BMC Bioinformatics* **7**, 109 (2006).
22. Hu, Z. *et al.* VisANT 3.5: multi-scale network visualization, analysis and inference based on the gene ontology. *Nucleic Acids Res.* **37** (web server issue), W115–W121 (2009).
23. McGuffin, M.J. & Jurisica, I. Interaction techniques for selecting and manipulating subgraphs in network visualizations. *IEEE Trans. Vis. Comput. Graph.* **15**, 937–944 (2009).
24. Prinz, S. *et al.* Control of yeast filamentous-form growth by modules in an integrated molecular network. *Genome Res.* **14**, 380–390 (2004).
25. Hu, Z. *et al.* Towards zoomable multidimensional maps of the cell. *Nat. Biotechnol.* **25**, 547–554 (2007).

26. Barsky, A., Munzner, T., Gardy, J. & Kincaid, R. Cerebral: visualizing multiple experimental conditions on a graph with biological context. *IEEE Trans. Vis. Comput. Graph.* **14**, 1253–1260 (2008).
27. Genc, B. & Dogrusoz, U. A layout algorithm for signaling pathways. *Inf. Sci.* **176**, 135–149 (2006).
28. de Lichtenberg, U., Jensen, L.J., Brunak, S. & Bork, P. Dynamic complex formation during the yeast cell cycle. *Science* **307**, 724–727 (2005).
29. Kanehisa, M. *et al.* KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* **36** (database issue), D480–D484 (2008).
30. Matthews, L. *et al.* Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.* **37** (database issue), D619–D622 (2009).
31. Kuhn, M. *et al.* STITCH 2: an interaction network database for small molecules and proteins. *Nucleic Acids Res.* **38** (database issue), D552–D556 (2010).
32. Quackenbush, J. Computational analysis of microarray data. *Nat. Rev. Genet.* **2**, 418–427 (2001).
33. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2009).
34. Bantscheff, M., Schirle, M., Sweetman, G., Rick, J. & Kuster, B. Quantitative mass spectrometry in proteomics: a critical review. *Anal. Bioanal. Chem.* **389**, 1017–1031 (2007).
35. Gstaiger, M. & Aebersold, R. Applying mass spectrometry-based proteomics to genetics, genomics and network biology. *Nat. Rev. Genet.* **10**, 617–627 (2009).
36. Westenberg, M.A., van Hijum, S.A.F.T., Kuipers, O.P. & Roerdink, J.B.T.M. Visualizing genome expression and regulatory network dynamics in genomic and metabolic context. *Comput. Graph. Forum* **27**, 887–894 (2008).
37. Freeman, T.C. *et al.* Construction, visualisation, and clustering of transcription networks from microarray expression data. *PLOS Comput. Biol.* **3**, e206 (2007).
38. Saraiya, P., Lee, P. & North, C. in *IEEE Symp. Information Visualization (InfoVis 2005)* 225–232 (2005).
39. Tufte, E.R. *The Visual Display of Quantitative Information* 2nd edn. (Graphics Press, Cheshire, Connecticut, USA, 2001).
40. Salomonis, N. *et al.* GenMAPP 2: new features and resources for pathway analysis. *BMC Bioinformatics* **8**, 217 (2007).
41. Neuweber, H. *et al.* Visualizing post genomics data-sets on customized pathway maps by ProMeTra – aeration-dependent gene expression and metabolism of *Corynebacterium glutamicum* as an example. *BMC Syst. Biol.* **3**, 82 (2009).
42. Kincaid, R., Kuchinsky, A. & Creech, M. VistaClara: an expression browser plug-in for Cytoscape. *Bioinformatics* **24**, 2112–2114 (2008).
43. Lee, H.K., Hsu, A.K., Sajdak, J., Qin, J. & Pavlidis, P. Coexpression analysis of human genes across many microarray data sets. *Genome Res.* **14**, 1085–1094 (2004).
44. Dunn, W.B. & Ellis, D.I. Metabolomics: current analytical platforms and methodologies. *TrAC Trends Anal. Chem.* **24**, 285–294 (2005).
45. Scholz, M. & Fiehn, O. Setup X – a public study design database for metabolomic projects. *Pac. Symp. Biocomput.* 169–180, doi:10.1142/9789812772435\_0017 (2007).
46. Thimm, O. *et al.* MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J.* **37**, 914–939 (2004).
47. Paley, S.M. & Karp, P.D. The Pathway Tools cellular overview diagram and omics viewer. *Nucleic Acids Res.* **34**, 3771–3778 (2006).
48. Tokimatsu, T. *et al.* KaPPA-view: a web-based analysis tool for integration of transcript and metabolite data on plant metabolic pathway maps. *Plant Physiol.* **138**, 1289–1300 (2005).
49. Mlecnik, B. *et al.* PathwayExplorer: web service for visualizing high-throughput expression data on biological pathways. *Nucleic Acids Res.* **33** (web server issue), W633–W637 (2005).
50. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* **36** (database issue), D623–D631 (2008).
51. Funahashi, A. *et al.* CellDesigner 3.5: a versatile modeling tool for biochemical networks. *Proc. IEEE* **96**, 1254–1265 (2008).
52. Sauro, H.M. *et al.* Next generation simulation tools: the Systems Biology Workbench and BioSPICE integration. *OMICS* **7**, 355–372 (2003).
53. Demir, E. *et al.* PATIKA: an integrated visual environment for collaborative construction and analysis of cellular pathways. *Bioinformatics* **18**, 996–1003 (2002).
54. Schreiber, F., Dwyer, T., Marriott, K. & Wybrow, M. A generic algorithm for layout of biological networks. *BMC Bioinformatics* **10**, 375 (2009).
55. Thomas, J.J. & Cook, K.A. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. (National Visual Analytics Center & IEEE, Richland, Washington, USA, 2005).



56. Saraiya, P., North, C. & Duca, K. An Insight-based methodology for evaluating bioinformatics visualizations. *IEEE Trans. Vis. Comput. Graph.* **11**, 443–456 (2005).
57. Dwyer, T., Koren, Y. & Marriott, K. IPSEP-COLA: an incremental procedure for separation constraint layout of graphs. *IEEE Trans. Vis. Comput. Graph.* **12**, 821–828 (2006).
58. Dwyer, T. *et al.* Exploration of networks using overview+detail with constraint-based cooperative layout. *IEEE Trans. Vis. Comput. Graph.* **14**, 1293–1300 (2008).
59. Viégas, F.B., Wattenberg, M., van Ham, F., Kriss, J. & McKeon, M. ManyEyes: a site for visualization at internet scale. *IEEE Trans. Vis. Comput. Graph.* **13**, 1121–1128 (2007).
60. Heer, J., Viégas, F.B. & Wattenberg, M. Voyagers and voyeurs: supporting asynchronous collaborative information visualization. in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'07)* 1029–1038 (ACM, New York, 2007).
61. Heer, J. & Agrawala, M. Design considerations for collaborative visual analytics. *Inf. Vis.* **7**, 49–62 (2008).
62. Pico, A.R. *et al.* WikiPathways: pathway editing for the people. *PLoS Biol.* **6**, e184 (2008).
63. Pavlopoulos, G.A. *et al.* Arena3D: visualization of biological networks in 3D. *BMC Syst. Biol.* **2**, 104 (2008).
64. Ball, R. & North, C. Realizing embodied interaction for visual analytics through large displays. *Comput. Graph.* **31**, 380–400 (2007).
65. Berman, H., Henrick, K. & Nakamura, H. Announcing the worldwide Protein Data Bank. *Nat. Struct. Biol.* **10**, 980 (2003).
66. Hermjakob, H. *et al.* The HUPPO PSI's molecular interaction format—a community standard for the representation of protein interaction data. *Nat. Biotechnol.* **22**, 177–183 (2004).
67. Hucka, M. *et al.* The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* **19**, 524–531 (2003).
68. Lloyd, C.M., Halstead, M.D. & Nielsen, P.F. CellML: its future, present and past. *Prog. Biophys. Mol. Biol.* **85**, 433–450 (2004).
69. Le Novère, N. *et al.* The Systems Biology Graphical Notation. *Nat. Biotechnol.* **27**, 735–741 (2009).
- This publication marks the first serious attempt to create a community standard for a graphical notation to represent networks in systems biology.**
70. Walter, T. *et al.* Visualization of image data from cells to organisms. *Nat. Methods* **7**, S26–S40 (2010).
71. O'Donoghue, S.I. *et al.* Visualization of macromolecular structures. *Nat. Methods* **7**, S42–S55 (2010).
72. Nielsen, C.B., Cantor, M., Dubchak, I., Gordon, D. & Wang, T. Visualizing genomes: techniques and challenges. *Nat. Methods* **7**, S5–S15 (2010).
73. Procter, J.B. *et al.* Visualization of multiple alignments, phylogenies and gene family evolution. *Nat. Methods* **7**, S16–S25 (2010).
74. Burrage, K., Hood, L. & Ragan, M.A. Advanced computing for systems biology. *Brief. Bioinform.* **7**, 390–398 (2006).
75. Awad, I.A., Rees, C.A., Hernandez-Boussard, T., Ball, C.A. & Sherlock, G. Caryoscope: an open source Java application for viewing microarray data in a genomic context. *BMC Bioinformatics* **5**, 151 (2004).
76. Lau, C. *et al.* Exploration and visualization of gene expression with neuroanatomy in the adult mouse brain. *BMC Bioinformatics* **9**, 153 (2008).
77. Weber, G.H. *et al.* Visual exploration of three-dimensional gene expression using physical views and linked abstract views. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **6**, 296–309 (2009).
78. Barsky, A., Gardy, J.L., Hancock, R.E. & Munzner, T. Cerebral: a Cytoscape plugin for layout of and interaction with biological networks using subcellular localization annotation. *Bioinformatics* **23**, 1040–1042 (2007).
79. Mookherjee, N. *et al.* Modulation of the TLR-mediated inflammatory response by the endogenous human host defense peptide LL-37. *J. Immunol.* **176**, 2455–2464 (2006).
80. Spellman, P.T. *et al.* Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* **9**, 3273–3297 (1998).
81. Kuntzer, J. *et al.* BNDB – the Biochemical Network Database. *BMC Bioinformatics* **8**, 367 (2007).
82. Baitaluk, M., Sedova, M., Ray, A. & Gupta, A. BiologicalNetworks: visualization and analysis tool for systems biology. *Nucleic Acids Res.* **34** (web server issue), W466–W471 (2006).
83. Cline, M.S. *et al.* Integration of biological networks and gene expression data using Cytoscape. *Nat. Protoc.* **2**, 2366–2382 (2007).
84. Hooper, S.D. & Bork, P. Medusa: a simple tool for interaction graph analysis. *Bioinformatics* **21**, 4432–4433 (2005).
85. Kao, H.L. & Gunsalus, K.C. Browsing multidimensional molecular networks with the generic network browser (N-Browse). *Curr. Protoc. Bioinformatics* **9**, 9.11.1–9.11.21 (2008).
86. Brown, K.R. *et al.* NAViGaTOR: Network Analysis, Visualization and Graphing Toronto. *Bioinformatics* **25**, 3327–3329 (2009).
87. Kohler, J. *et al.* Graph-based analysis and visualization of experimental results with ONDEX. *Bioinformatics* **22**, 1383–1390 (2006).
88. Breitkreutz, B.J., Stark, C. & Tyers, M. Osprey: a network visualization system. *Genome Biol.* **4**, R22 (2003).
89. Batagelj, V. & Mrvar, A. Pajek – Program for large network analysis. *Connections* **21**, 47–57 (1998).
90. Forman, J.J., Clemons, P.A., Schreiber, S.L. & Haggarty, S.J. SpectralNET—an application for spectral graph analysis and visualization. *BMC Bioinformatics* **6**, 260 (2005).
91. Auber, D. A huge graph visualization framework. in *Graph Drawing Software* (eds. Mutzel, P. & Jünger, M.) 105–126 (Springer, Heidelberg, Germany, 2004).
92. Zinovyev, A., Viara, E., Calzone, L. & Barillot, E. BiNoM: a Cytoscape plugin for manipulating and analyzing biological networks. *Bioinformatics* **24**, 876–877 (2008).
93. Huttenhower, C., Mehmood, S.O. & Troyanskaya, O.G. Graphlet: interactive exploration of large, dense graphs. *BMC Bioinformatics* **10**, 417 (2009).
94. Longabaugh, W.J., Davidson, E.H. & Bolouri, H. Visualization, documentation, analysis, and communication of large-scale gene regulatory networks. *Biochim. Biophys. Acta* **1789**, 363–374 (2009).
95. Streit, M., Lex, A., Kalkusch, M., Zatloukal, K. & Schmalstieg, D. Caleydo: connecting pathways and gene expression. *Bioinformatics* **25**, 2760–2761 (2009).
96. Okuda, S. *et al.* KEGG atlas mapping for global analysis of metabolic pathways. *Nucleic Acids Res.* **36** (web server issue), W423–W426 (2008).
97. van Iersel, M.P. *et al.* Presenting and exploring biological pathways with PathVisio. *BMC Bioinformatics* **9**, 399 (2008).
98. Holford, M., Li, N., Nadkarni, P. & Zhao, H. VitaPad: visualization tools for the analysis of pathway data. *Bioinformatics* **21**, 1596–1602 (2005).
99. Chung, H. J., Kim, M., Park, C. H., Kim, J., and Kim, J. H. ArrayXPath: mapping and visualizing microarray gene-expression data with integrated biological pathway resources using scalable vector graphics. *Nucleic Acids Res.* **32** (web server issue), W460–W464 (2004).
100. Weniger, M., Engelmann, J.C. & Schultz, J. Genome Expression Pathway Analysis Tool—analysis and visualization of microarray gene expression data under genomic, proteomic and metabolic context. *BMC Bioinformatics* **8**, 179 (2007).
101. Letunic, I., Yamada, T., Kanehisa, M. & Bork, P. iPath: interactive exploration of biochemical pathways and networks. *Trends Biochem. Sci.* **33**, 101–103 (2008).
102. Karp, P.D., Paley, S. & Romero, P. The Pathway Tools software. *Bioinformatics* **18** (suppl. 1), S225–S232 (2002).
103. Dogrusoz, U. *et al.* PATIKAwab: a Web interface for analyzing biological pathways through advanced querying and visualization. *Bioinformatics* **22**, 374–375 (2006).
104. Santamaria, R., Theron, R. & Quintales, L. BicOverlapper: a tool for bicluster visualization. *Bioinformatics* **24**, 1212–1213 (2008).
105. Goncalves, J.P., Madeira, S.C. & Oliveira, A.L. BiGGESTs: integrated environment for biclustering analysis of time series gene expression data. *BMC Res Notes* **2**, 124 (2009).
106. Shamir, R. *et al.* EXPANDER—an integrative program suite for microarray data analysis. *BMC Bioinformatics* **6**, 232 (2005).
107. Sturn, A., Quackenbush, J. & Trajanoski, Z. Genesis: cluster analysis of microarray data. *Bioinformatics* **18**, 207–208 (2002).
108. Hibbs, M.A., Dirksen, N.C., Li, K. & Troyanskaya, O.G. Visualization methods for statistical analysis of microarray clusters. *BMC Bioinformatics* **6**, 115 (2005).
109. Seo, J. & Shneiderman, B. Interactively exploring hierarchical clustering results. *Computer* **35**, 80–86 (2002).
110. Saldanha, A.J. Java Treeview—extensible visualization of microarray data. *Bioinformatics* **20**, 3246–3248 (2004).
111. Dietzsch, J., Gehlenborg, N. & Nieselt, K. Mayday – a microarray data analysis workbench. *Bioinformatics* **22**, 1010–1012 (2006).
112. Gehlenborg, N., Dietzsch, J. & Nieselt, K. A framework for visualization of microarray data and integrated meta information. *Inf. Vis.* **4**, 164–175 (2005).
113. Saeed, A.I. *et al.* TM4: a free, open-source system for microarray data management and analysis. *Biotechniques* **34**, 374–378 (2003).
- One of the first and still one of the most commonly used applications for the management, analysis and visualization of microarray data.**
114. Hochheiser, H., Baehrecke, E.H., Mount, S.M. & Shneiderman, B. Dynamic querying for pattern identification in microarray and genomic data. *Proc. IEEE Multimedia and Expo Int. Conf.* **3**, 453–456 (2003).



115. Kapushesky, M. *et al.* Expression Profiler: next generation—an online platform for analysis of microarray data. *Nucleic Acids Res.* **32** (web server issue), W465–W470 (2004).
  116. Reich, M. *et al.* GenePattern 2.0. *Nat. Genet.* **38**, 500–501 (2006).
  117. Quackenbush, J. Microarray data normalization and transformation. *Nat. Genet.* **32** (suppl.), 496–501 (2002).
  118. Brettschneider, J., Collin, F., Bolstad, B.M. & Speed, T.P. Quality assessment for short oligonucleotide microarray data. *Technometrics* **50**, 241–264 (2008).
  119. Kauffmann, A., Gentleman, R. & Huber, W. ArrayQualityMetrics—a bioconductor package for quality assessment of microarray data. *Bioinformatics* **25**, 415–416 (2009).
  120. Morgan, M. *et al.* ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics* **25**, 2607–2608 (2009).
  121. Robinson, M.D., McCarthy, D.J. & Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
  122. Smyth, G.K., Yang, Y.H. & Speed, T. Statistical issues in cDNA microarray data analysis. *Methods Mol. Biol.* **224**, 111–136 (2003).
  123. Nesvizhskii, A.I., Vitek, O. & Aebersold, R. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Methods* **4**, 787–797 (2007).
  124. Perkins, D.N., Pappin, D.J., Creasy, D.M. & Cottrell, J.S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **20**, 3551–3567 (1999).
  125. Li, X.J. *et al.* A tool to visualize and evaluate data obtained by liquid chromatography-electrospray ionization-mass spectrometry. *Anal. Chem.* **76**, 3856–3860 (2004).
  126. Sturm, M. & Kohlbacher, O. TOPPView: an open-source viewer for mass spectrometry data. *J. Proteome Res.* **8**, 3760–3763 (2009).
  127. Kopka, J. Current challenges and developments in GC-MS based metabolite profiling technology. *J. Biotechnol.* **124**, 312–322 (2006).
  128. Broeckling, C.D. *et al.* Metabolic profiling of *Medicago truncatula* cell cultures reveals the effects of biotic and abiotic elicitors on metabolism. *J. Exp. Bot.* **56**, 323–336 (2005).
  129. Beckonert, O. *et al.* Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nat. Protoc.* **2**, 2692–2703 (2007).
  130. Lindon, J.C. & Nicholson, J.K. Spectroscopic and statistical techniques for information recovery in metabonomics and metabolomics. *Annu. Rev. Anal. Chem.* **1**, 45–69 (2008).
  131. Xia, J., Bjorndahl, T.C., Tang, P. & Wishart, D.S. MetaboMiner—semi-automated identification of metabolites from 2D NMR spectra of complex biofluids. *BMC Bioinformatics* **9**, 507 (2008).
  132. Hotelling, H. Analysis of complex statistical variables into principal components. *J. Educ. Psychol.* **24**, 417–441 (1933).
  133. Kruskal, J. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* **29**, 1–26 (1964).
  134. Venna, J. & Kaski, S. Comparison of visualization methods for an atlas of gene expression data sets. *Inf. Vis.* **6**, 139–154 (2007).
  135. Inselberg, A. The plane with parallel coordinates. *Vis. Comput.* **1**, 69–91 (1985).
  136. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95**, 14863–14868 (1998).
- Milestone publication that introduced the heat map visualization to the field of transcriptomics and has been cited several thousand times.**
137. Wilkinson, L. & Friendly, M. The history of the cluster heat map. *Am. Stat.* **63**, 179–184 (2009).
  138. Weinstein, J.N. Biochemistry. a postgenomic visual icon. *Science* **319**, 1772–1773 (2008).