Visualization of Biomedical Data

Corresponding author: Seán I. O'Donoghue; email: sean@odonoghuelab.org

- Data61, Commonwealth Scientific and Industrial Research Organisation (CSIRO), Eveleigh NSW 2015, Australia
- Genomics and Epigenetics Division, Garvan Institute of Medical Research, Sydney NSW 2010, Australia
- School of Biotechnology and Biomolecular Sciences, UNSW, Kensington NSW 2033, Australia

Benedetta Frida Baldi; email: b.baldi@garvan.org.au

- Genomics and Epigenetics Division, Garvan Institute of Medical Research, Sydney NSW 2010, Australia

Susan J Clark; email: s.clark@garvan.org.au

- Genomics and Epigenetics Division, Garvan Institute of Medical Research, Sydney NSW 2010, Australia

Aaron E. Darling; email: aaron.darling@uts.edu.au

- The ithree institute, University of Technology Sydney, Ultimo NSW 2007, Australia

James M. Hogan; email: j.hogan@qut.edu.au

- School of Electrical Engineering and Computer Science, Queensland University of Technology, Brisbane QLD, 4000, Australia

Sandeep Kaur; email: sandeep.kaur@unsw.edu.au

- School of Computer Science and Engineering, UNSW, Kensington NSW 2033, Australia

Lena Maier-Hein; email: l.maier-hein@dkfz-heidelberg.de

- Div. Computer Assisted Medical Interventions (CAMI), German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany

Davis J. McCarthy; email: davis@ebi.ac.uk

- European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, CB10 1SD, Hinxton, Cambridge, UK
- St. Vincent's Institute of Medical Research, Fitzroy VIC 3065, Australia

William J. Moore; email: w.moore@dundee.ac.uk

- School of Life Sciences, University of Dundee, Scotland, DD1 5EH, UK

Esther Stenau; email: e.stenau@dkfz-heidelberg.de

- Div. Computer Assisted Medical Interventions (CAMI), German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany

Jason R. Swedlow; email: j.r.swedlow@dundee.ac.uk

- School of Life Sciences, University of Dundee, Scotland, DD1 5EH, UK

Jenny Vuong; email: vuong.jenny@gmail.com

- Data61, Commonwealth Scientific and Industrial Research Organisation (CSIRO), Eveleigh NSW 2015, Australia

James B. Procter; email: j.procter@dundee.ac.uk

- School of Life Sciences, University of Dundee, Scotland, DD1 5EH, UK

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26896v1 | CC BY 4.0 Open Access | rec: 26 Apr 2018, publ: 26 Apr 2018

Keywords

Data visualization; Multivariate data; Molecular biology; Cell biology; Tissue imaging; Metagenomics

Abstract

The rapid increase in volume and complexity of biomedical data requires changes in research, communication, training, and clinical practices. This includes learning how to effectively integrate automated analysis with high-data-density visualizations that clearly express complex phenomena. In this review, we summarize key principles and resources from data visualization research that address this difficult challenge. We then survey how visualization is being used in a selection of emerging biomedical research areas, including: 3D genomics, single-cell RNA-seq, the protein structure universe, phosphoproteomics, augmented-reality surgery, and metagenomics. While specific areas need highly tailored visualization tools, there are common visualization challenges that can be addressed with general methods and strategies. Unfortunately, poor visualization practices are also common; however, there are good prospects for improvements and innovations that will revolutionize how we see and think about our data. We outline initiatives aimed at fostering these improvements via better tools, peer-to-peer learning, and interdisciplinary collaboration with computer scientists, science communicators, and graphic designers.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26896v1 | CC BY 4.0 Open Access | rec: 26 Apr 2018, publ: 26 Apr 2018

INTRODUCTION

The launch of this annual review journal is driven by the rapid increase in volume and complexity of biomedical data, which requires changes in research, communication, training, and clinical practices (1). Without these changes, many biomedical discoveries will remain buried in data already collected, and many misdiagnoses (now estimated at ~10-30% of all diagnoses) will remain unrecognized (2, 3), contributing to a major cause of death (4).

These changes in practice will include the development and adoption of new, automated analysis methods (e.g., clustering, modeling, and 'deep' machine learning). However, while necessary (**Figure 1*a***), automated analysis is not sufficient: as demonstrated by 'Anscombe's quartet' (**Figure 1*b***), to find the truth, we need to visually inspect all relevant data and analyses combined[1]. Driven by this realization, **data visualization** has been a major research focus in computer science for decades (**Sidebar: Data Visualization, SciVis, and InfoVis**), yielding many resources that could accelerate discovery in biomedical research (5-10).

Unfortunately, relatively few scientists currently use these resources. This is evident, for example, from the widespread use of rainbow color maps (found in >50% of papers in a survey of research publications containing scientific visualizations, (11)). Whilst seemingly inconsequential, visualization research has shown that rainbow maps can obfuscate true data patterns and introduce visual artifacts (11). Sadly, many biomedical datasets (often difficult or expensive to acquire) are inspected using poor visualization methods, although better alternatives are known.

Similarly, visualization methods are underutilized by clinicians, contributing to misdiagnoses. About half of all diagnostic errors arise from faulty cognitive processing of data (2); many of these errors can be addressed by improving how data is visualized (12). This is especially true in fields of medicine, such as radiology, where the core data are intrinsically visual, and diagnosis depends largely on visual perception (3).

We believe that the underuse of visualization methods has arisen largely because of the following misconceptions we often encounter:

Misconception 1: 'The goal of data visualization is to impress'. We sometimes think of data visualization as purely aesthetic; adding an optional 'wow' factor not present in the data itself. This can be true when creating artwork (e.g., a cover figure); but the role of data visualization in research is almost exactly the opposite: it is a necessary step, aimed at clearly revealing patterns in data.

Misconception 2: 'Data visualization is easy'. Well-designed visualizations can be so easy to understand and use that we are misled into thinking they must have been easy to create. However: "most graphs are simple, but their invention was neither simple nor obvious – the idea did not occur to the Greeks or Romans, nor even to the great 17th Century mathematician-experimenters such as Newton and Leibniz." (Lewandowsky & Spence, 13).

Misconception 3: 'Studying data visualization is unnecessary'. Underestimating the difficulty of data visualization can lead us to overestimate our current skills and conclude we would gain little benefit from investing time, effort, or money in training or study.

[1] "The eye of a Master, will do more Work than his Hand", Benjamin Franklin (1744). *Poor Richard*.

Misconception 4: 'Visualization is just a synonym for imaging'. In the life sciences, 'visualization' is often used as a synonym for imaging experiments. In computer science, 'data visualization' has a much broader meaning: in addition to imaging data, it encompasses abstract data, interactive analysis, design, as well as visual and cognitive abilities – and its purpose is insight, not pictures (10).

Below, we outline some key data visualization principles: many are straightforward, helping to create better visualizations and avoid common mistakes (**Sidebar: Avoiding Common Visualization Mistakes**). We then survey a selection of emerging biomedical research areas where visualization is playing a key role. Finally, we discuss how visualization can enhance biomedical communication, and offer perspectives for improving the global standard of data visualization in biomedical research.

DATA VISUALIZATION PRINCIPLES

Data Volume

How can data visualization help us deal with the increasingly large volume of biomedical datasets?

One straightforward answer is: get more pixels. Our visual system has extraordinary capacity, and can manage much more information than is presented in many of the scientific visualizations we currently create. Information transfer speed from the eye to the brain is about 10 megabits/second, similar to wired internet (14), and well-encoded visual patterns can be recognized within 250 ms ('pre-attentively', 15). To use more of this capacity, larger, higher resolution displays can help: for example, connecting a laptop to a 4K display can be a cost-effective way to see more detail, and to improve navigation and work efficiency (16). However, larger displays become increasingly less cost-effective, and often have impractical user interface controls. In addition, scaling up a visualization can make global patterns (e.g., correlation) harder to perceive (17) – clearly, there is an optimal size range for visualizations.

Therefore, a second answer is: create visualizations with greater **data density**. Here again, our visual system has greater capacity than we typically use: aided by redundancy and context, the eye can resolve features to 0.1 mm (5). Thus, many visualization researchers advocate creating visualizations with high data density - often a practical requirement in high impact journals, due to space limitations (5). Creating compact visualizations requires carefully selecting **visual channels** that encode data with high **visual effectiveness**: fortunately, visualization research provides clear guidelines for this (**Figure 2**). Also required is time and effort to learn specialist tools (**Table 1**) that provide precise controls necessary for visualizing large data volumes with high data density. However, there are limits: as data density increases, visualizations can require specific, targeted strategies to remain effective (18).

Data Complexity

In many ways, 'big data' volume is a small problem in the life sciences: far more challenging is the complexity of our data, which is often **multivariate**, multi-scale, highly interconnected, and dependent on very specific conditions.

Here, a common strategy is to use analytical methods to reduce dimensionality (e.g., clustering, principal component analysis). However, Anscombe's quartet (**Figure 1*b***) reminds us we need to visually inspect all relevant data before we draw conclusions from a simplified subset – and very often we require many more than two variables to express the complex phenomena studied in biomedical datasets.

Remarkably, multivariate data of any dimensionality can be visualized in two dimensions (2D) without loss of information, using a range of generic methods (**Figure 3*a-d***) and tools (**Table 1**). However, multidimensional data patterns are often scrambled and hard to recognize or interpret when encoded into 2D (19).

Data Integration & Tailored Visualizations

Clearly revealing these multidimensional patterns usually requires carefully **tailored visualizations** that use very specific data integration strategies. **Figure 3*e*** reproduces an exemplary tailored visualization created by Charles Minard in 1869, showing Napoleon's disastrous Russian campaign (5). To learn how to make better visualizations, it is worthwhile studying in detail (guided by **Figure 2**) how this complex, multivariate data story has been communicated so clearly and concisely.

While generic visualization methods (**Figure 3*a-d***) and tools are frequently used in biomedical research, tailored visualizations are the mainstay. Creating them requires three inter-related steps:

1. Identify the necessary complexity: this is the subset of data that can **visually express** all - and only - information of most relevance to the phenomena studied. This usually means excluding data (e.g., in **Figure 3*e***, temperature is shown only when it caused significant losses), or showing only derived features (e.g., principal components).
2. Identify necessary supporting context: this often means adding related information not part of the initial dataset (e.g., in **Figure 3*e***, geographic features help interpret the dataset).
3. Invent a visual strategy that coherently integrates these data, using effective visual encoding and conventions familiar to peers.

A successful tailored visualization arranges all relevant data into a compact, immediately accessible, 2D view. This facilitates **spatial reasoning**, which reduces the cognitive load needed to 'read' a visualization, and gain insight from data (20). By using familiar or intuitive visual conventions, successful visualization strategies also reduce the cognitive load needed when first learning to how to read them.

Fortunately, for many of the data challenges faced in biomedical research (21-25), tailored visualization methods have already been invented and implemented into working tools. These tools often use interactivity (9, 26) to facilitate combined exploration and integration of raw data, data derived via analysis, and additional supporting evidence.

However, cutting-edge research often requires us to invent novel, tailored visualizations. This difficult task can be aided by drawing ideas and inspiration from visualization resources (see **Table 1** and annotated references, (1, 5-10, 27-29)), and also by learning from peers facing similar challenges. Thus, the next section surveys tailored visualizations being used to reveal new insights across a broad range of biomedical research areas.

VISUALIZATION FOR DISCOVERY

Genomics & Epigenetics

We begin this survey in the field of genomics and epigenetics, where rapid advances in DNA sequencing technologies are generating a flood of data. These data are not just limited to raw DNA sequences, but include an increasing spectrum of additional information that can be obtained genome-wide: for example, single-molecule bisulphite sequencing can determine the methylation state of every cytosine base; chromatin immunoprecipitation sequencing (ChIP-seq) can map protein-DNA interactions; and Assay for Transposase-Accessible Chromatin (ATAC-seq) can map chromatin accessibility. Together, this flood of data contains unprecedented detail into the molecular structure, regulation, and function of whole organisms – but in a condensed, fragmented, and encoded form.

Visualization is widely used to help unravel this information: a very common task – and often a rate-limiting step - is manual browsing of features to gain insight into function (21). The linear organization of chromosomes (**Figure 4*a***) provides a natural visual layout, allowing many genomic features to be positioned on a common horizontal scale (**Figure 4*b***) – thus using the most effective visual channel (**Figure 2**).

Visualization challenges. A core challenge is multiscale navigation - both horizontally (across hundreds of millions of base pairs) and vertically (since regions can contain thousands of overlapping genomic features). Current genome browsers (21, 30) address this fairly well, using the general strategy of **overview first, details upon demand (overview/details)**; this strategy is implemented using feature clustering methods (e.g., ChromHMM (31)) and user-interface controls to help users find and explore specific genomic regions and features of interest, while maintaining awareness of overall chromosomal location and context (**Figure 4*b***).

As new genomic technologies (e.g., single-cell or single-molecule DNA sequencing) continue to produce data of rapidly increasing volume and complexity, further innovations are needed in visualization methods. This includes improvements in multiscale navigation, error and uncertainty visualization (e.g., arising from base calling, assembly, and finishing, 21), variant analysis (32), and in managing *de novo* assemblies for organisms where reference genomes are not available. In addition, novel tailored visualizations need to be developed to address a wide range of important, yet very specific biomedical topics, such as genomic rearrangements in cancer (33).

Emerging frontiers. An exciting frontier in genomics is the study of the three-dimensional (3D) spatial organization of chromosomes. An accurate, atomic-scale model of the genome is a grand challenge that may someday be plausible thanks to recently developed experimental techniques – primarily, Hi-C (34) - that can determine spatial chromatin contacts between pairs of genomic regions. These methods have low resolution and high false-positive rates, so cannot yet determine accurate 3D models for chromosomes (35). Nonetheless, Hi-C data can give new insights - but interpreting these datasets is difficult (36). Thus, tailored visualization methods are being developed, currently based around three **alternative views**. In one, Hi-C data are shown as a contact matrix (**Figure 4*c***), allowing for high data density and a clear overview (37), but making it difficult to overlay other genomic features. A pyramidal layout (**Figure 4*d***) addresses this issue (38), but makes it harder to see contacting regions. A circular layout (**Figure 4*e***) is more compact (39), and using arcs to show contacts is a more effective visual encoding (40); but this does not allow the same data density, hence only major contacts (calculated via clustering) are shown. Such tradeoffs between alternative views of multidimensional data are common, with each viewpoint providing different insights.

These and other advances in genomics are gradually unravelling the remaining mysteries of genomic function. Currently, however, we still lack an understanding of many core process, such as exactly what gets transcribed, and when and how this is controlled in different cell types.

RNA Biology

RNA molecules play a leading role in biological systems, acting as messengers, sensors, and forming the ribosome, one of the most ancient molecular machines. Many researchers now focus on unlocking the secrets of the RNA world; meanwhile, the measurement of RNA transcript abundances has become the workhorse of modern biology. First accomplished with microarray experiments (41), accurate measurements of the abundance of transcripts in biological samples and single cells (42) are now taken with RNA sequencing (RNA-seq) (43).

Visualization challenges. Interpreting the high-dimensional datasets from RNA-seq experiments remains challenging. After careful experimental design and statistical analysis (44), gene expression values judged to be significant are commonly presented as clustered heat maps (45), a technique that has dominated since the first microarray experiments (41). However, optical illusions in these visualizations make it difficult to judge magnitude for individual values, or fold-changes between pairs of values (**Sidebar: Optical Illusions Caused by Ground Subtraction**). As the number of rows and column increase or cell size is reduced, these effects become worse, making it impractical to display all significant results as one very large heat map. Further problems arise because the rows and columns of clustered heat maps are usually ordered to group associated genes and conditions, and so highlight regulatory effects (46). Inevitably, values for genes and conditions without significant association will be placed next to one another, which exacerbates perceptual problems. Separating unrelated rows and columns (**Figure 5*a***) can help, but does not fundamentally address these difficulties (47), particularly for genes that cluster poorly. In such cases, there may be insufficient data to resolve those genes' regulatory networks as a one-dimensional ordering, so it is important that the degree and support for relationships inferred from clustering are also shown. The addition of tree graphs, however, further constrains the size of heat map that can be displayed without issue[2], so we suggest only the most informative subset of genes and conditions should be presented in this way.

Emerging frontiers. Single-cell RNA-seq experiments ('scRNA-Seq') are a revolutionary new technology that can reveal key events in differentiation normally masked in bulk RNA-Seq experiments, thus providing deep insight into the behavior of cells and tissues. These data are typically visualized using dimensionality reduction methods that allow gene expression vectors to be projected onto two-dimensional scatter plots (**Figure 5*b-d***). scRNA-Seq data allow the sequence of these events to be reconstructed – commonly referred to as cell pseudo-time (48). Clustering and dimensionality reduction heuristics (48) allow pseudo-time to be inferred, visualized, and quantitatively analyzed (**Figure 5*d***).

Single-cell transcriptomics measurements will soon become possible at the whole-organism level (49), and we will undoubtedly require more effective methods for interpreting these data. However, the measures of abundance obtained from these experiments are only markers that indicate which parts of an organism's genome are active. In order to understand the biological role each gene plays, we must look beyond sequencing data; in fact, much of our current understanding has come from studying the molecular structure of RNA transcripts, and the proteins they encode.

[2] We recommend rectangular heat map cells of no less than 6 mm, separated by 1.5 mm, and overlaid on a neutral background (white, black, or a color that does not contrast with those employed in the heat map).

Protein Structures

Protein structural biology aims to provide a detailed understanding of life's molecular machinery. Thanks to decades of research worldwide, we now have 3D molecular structures (at or near atomic-resolution) for ~40,000 proteins. By viewing these structures, researchers can gain insight into precise molecular mechanisms underlying many of the biochemical processes occurring within living cells. Remarkably, almost all these structures are collected in a single, exemplary database (https://wwpdb.org, 50) (**Sidebar: An Exemplary Biomedical Databank**); this has helped drive innovation in molecular graphics, which has outpaced visualization advances in many other areas of biomedical science (23).

Visualization challenges. Visualization is integral to structure determination and validation (23), as well as for gaining insight into protein function (e.g., with tools such as Chimera and others, 23, 51, 52). A core challenge is conveying the many different features of these large, complex datasets; this requires careful use of visualization principles (e.g., overview/details), judicious and minimal use of color, and visually expressive representations to highlight specific aspects of the data (**Figure 6*a-b***).

Another challenge is conveying the complex 3D shape of proteins. In special cases, shape can be communicated through specifically tailored 2D visualizations (**Figure 6*c***); but ultimately, protein structures need to be viewed in 3D. For this reason, structural biology has been an early-adopter of new visual techniques, starting with physical models (used in solving the first protein structures, 53), stereoscopic imaging (54), and **virtual reality (VR)** (55)[3]. This has continued, with techniques such as low-cost VR (e.g., VMD supports Oculus Rift, 56), very low-cost VR (e.g., Autodesk Molecule Viewer supports Google Cardboard, https://www.molviewer.com/), 3D printing (57), commodity interaction devices (e.g., Leap motion & Kinect, 58), **augmented reality (AR)** (59), crowdsourced evaluations (60), concepts from computer gaming (61), as well as emerging web technologies (e.g., WebGL; **Figure 6*a***).

Emerging frontiers. Protein structural biology is still far from complete, as many proteins still have little or no experimentally-determined structural information. To address this, high-throughput homology modeling is being used to systematically compare all known protein sequences against all experimentally determined structures, resulting in >100 million model structures (62). Allowing researchers to effectively explore and benefit from such large datasets requires carefully tailored visualization tools (62) that use the overview/detail strategy, as well as alternative views connected via **brushing and linking** (**Figure 6*d* & 6*e***). Homology modeling currently provides structural models for about half of the eukaryotic proteome (**Figure 6*e***). Interestingly, much of the remaining 'dark' proteome currently cannot be explained (e.g., **Figure 6*f***) – exploring this dark 'protein structure universe' is an important data science challenge (63) in which visualization is playing a key role (64).

High-throughput approaches are also being applied to molecular dynamics (65), generating increasingly large, complex trajectories; these data can give insights into key events (e.g., binding with ligands or other proteins). However, unearthing those insights is a still major challenge, requiring further innovations to create very specific, tailored visualization tools (e.g., 23, 51, 56).

[3] VR can focus undivided attention on a dataset; although powerful, usage is limited by inconvenience, discomfort, motion sickness, and other drawbacks. By contrast, AR has fewer drawbacks and looks likely to become widespread in biomedical research – and in normal life.

Rapid advances in cryo-electron microscopy (cryo-EM) are making accessible much larger structures and molecular assemblies than ever before (66); this has promoted improvements in methods for visual exploration of multiscale molecular data (http://ncbr.muni.cz/LiteMol).

Finally, high-throughput computing is also being used to integrate structural data in the construction of atomic-scale models of viruses, subcellular compartments, or even whole cells (67). The scale and complexity of these models requires the development of radically new visualization methods, bridging structural and systems biology (68).

Systems Biology

We have long speculated how biomolecules coordinate to perform cellular function (69) – and graph-based visualizations have been key to organizing our thoughts (24). An exemplar is the Roche metabolic pathway (**Figure 7*a***), initiated by Gerhard Michal in 1965 (http://biochemical-pathways.com/, 70); this manually tailored visualization shows thousands of metabolic reactions in a single, comprehensive view. Such pathway graphs have endured because they are visually expressive, showing causal flow and providing insight into molecular events underlying health and disease.

Visualization challenges. Over 4 billion biochemical reactions are currently known - and this number is rising rapidly (71). These data are typically visualized with specialized tools (e.g., Cytoscape (72) or Gephi (73)) that provide a range of automated layout methods, many based on force-directed algorithms (74), resulting in network graphs (**Figure 7*b***). A force-directed layout (also known as spring-embedding) can be useful for overviewing a dataset; however even small biological networks are often so interconnected that these graphs become overly cluttered (**Figure 7*b***). Force-directed layout is so common it has become something of a limiting paradigm, often used even when better strategies are available[4]. For example, when integrating connectivity with other data (e.g., time, subcellular location, etc.), the go-to strategy has been to overlay these data onto existing network layouts (24), thus reducing visual effectiveness, due to clutter. Often, it is better to change the layout entirely, using position (the most effective visual channel for quantitative data; **Figure 2**) to encode not just connectivity, but also biological context that helps in interpreting data. This approach is taken in a number of tailored visualization tools, such as Cerebral (75), which uses position to encode subcellular location, edge-bundling to reduce clutter (76), and small multiples for different conditions (77).

Visually expressive layouts (e.g., Cerebral) need to be tailored for each specific scenario - and there are vastly many distinct scenarios in systems biology, since reactions vary greatly with cell-type, timing, and molecular microenvironment (78). Thus, many tailored visualizations have already been developed (24, 30), and, unfortunately, systems biology data are fragmented across ~700 resources (http://pathguide.org, 71), which partly impedes progress in the field. To help manage the many complexities of these data, a wide range of visual techniques are used; for example, the overview/detail strategy is used to allow subgraphs to be collapsed or expanded upon demand, thus helping users more effectively explore large graphs. However, there remains considerable scope for using visualization principles to design new, automated layouts (79, 80), and to improve the computer-aided design of manual layouts. There is also considerable scope for improving how systems biology data are organized (An Exemplary Biomedical Databank).

[4] Over-reliance on familiar tools can lead to cognitive bias: paraphrasing a popular adage, if your only tool is spring-embedding, every dataset looks like a network graph.

Emerging frontiers. Mass spectrometry-based proteomics (81, 82) is a rapidly emerging technology that enables systematic measurement of proteome-wide posttranslational modifications, such as phosphorylation (termed 'phosphoproteomics'), in response to stimuli. These technologies are providing new insights into fundamental cellular processes and into diseases, which in term may lead to new therapeutic interventions. However, there is a price: increased complexity. Each phosphoproteomics experiment can track highly dynamic changes in over 10,000 different phosophosites in >5,000 proteins (82). In analyzing these datasets, it is common to first identify co-regulated phosophosites, based on clustering of time-profiles (**Figure 7*c***). Several tailored visualization strategies are being developed for exploring these clusters. In one, inspired by the cyclic journey in Minard's exemplary chart (**Figure 3*e***), the cascade of phosphorylation events is laid out as a cyclic journey through a cellular landscape (**Figure 7*d***). Proteins are represented as tracks and phospho-events are positioned by time and subcellular location – two key variables in these experiments. This layout facilitates spatial reasoning about causal relationships, helping researchers use these complex datasets to gain insight into cellular processes, such as insulin response (83) or mitosis (84).

These and other advances in molecular systems biology promise to revolutionize medicine. However, realizing this promise will require software platforms capable of bridging scales, from molecules to cells, tissues, and whole organisms – a formidable challenge in which visualization plays a central role (68).

Cellular & Tissue Imaging

Imaging remains the primary way that we observe biological systems. Quantitative imaging data are employed throughout the biomedical sciences as a basis for research, diagnosis, and therapy. Since van Leuwenhoek created the first microscope, biologists have observed the structure and behavior of cells, how they form tissues and develop into organisms. Most recently, technological advances in labeling, sample handling, and imaging have expanded microscopy's capabilities (85), and we are now able to image complex cellular assemblies, such as neurons, in 3D, or capture in real time the cellular processes that drive development (86). Imaging has also advanced in medicine, allowing detection and diagnosis of pathologies, and providing essential guidance before and during surgery.

Visualization challenges. Imaging data from can give quantitative and qualitative insights into morphology and function; however, this often requires extensive and sophisticated processing pipelines (87). An increasing array of tools are being developed to address this need, some of which now allow interactive visualization of raw images together with image-derived data - e.g., MITK (http://mitk.org/) and Slicer (http://www.slicer.org/). The primary way that we interpret these image-derived data is by encoding them as colors and annotations that are then overlaid onto the original images. This presents a challenge, because biological images are usually already very complex; thus, detail often needs to be removed from the original image – for example, by transforming the image's dynamic range or color space – before derived data can be overlaid, either for data exploration or for publication. Whilst many image processing and figure generation tools allow these operations, such manipulations may obscure critical details, and therefore need to be documented in a reproducible manner; there is broad consensus that the transformation of image data needs to be better reported in scientific publications (88). Recent advances (89) are beginning to address these issues, allowing interactive creation of figures for online publication (**Figure 8*a***) that link to original data (e.g., high content screening, time-lapse or histological whole slide imaging data) as well as metadata (related to experimental design, image acquisition, downstream analysis and interpretation) – thus, allowing subsequent re-analysis.

Emerging frontiers. An exciting frontier in medicine is the use of augmented reality (90) to enhance, for example, live video feeds used to guide surgery with pre-operative, diagnostic imaging data (**Figure 8*b***). This approach promises to reduce error and increase precision during surgery by providing real-time guidance about the location and physiological status of diseased tissue. Visual complexity is, once again, a major challenge; however, by integrating machine learning and semantic modeling approaches (91), surgeons today have already begun to use AR to see beyond the capabilities of normal visual perception.

There remains tremendous opportunity for improving the integration of advanced analytic tools with interactive visualization, creating new platform, such as the Allen Cell Explorer (http://www.allencell.org/). In the near future, such improvements are set to greatly advance our understanding of the composition, structure, and dynamics of normal and pathological cells and tissues, as well as the effectiveness and precision of medical interventions.

Populations & Ecosystems

Some of the most compelling questions in biology center on how our genome affects how we live and interact with other organisms and our environment, and how these interactions change over time. Tree graphs (**Figure 9*a***) are the primary way we visualize ancestral relationships between organisms. With sufficient data, trees can convey not only evolutionary distance but also the order in which different lineages may have evolved. Genomic sequence comparisons allow us to infer how species or individuals within a population differ from one another, but these contain localized features (e.g., SNPs and indels), copy number variants, and rearrangements that extend over millions of bases. Parallel coordinates (**Figure 9*b***) provide one way of viewing multiple alignments of closely related bacterial genomes. These visualizations highlight the changes in genomic regions or even individual genes as they undergo mutation, rearrangements, or horizontal gene transfer.

Visualization challenges. Phylogenetic tree and comparative genomic visualizations are relatively mature (21, 30). For metagenomics and population sequencing, however, neither representation will suffice. Here, interactive plots or static visualizations at multiple scales are needed to capture the breadth and depth of these data. Branch structures in phylogenetic trees quickly become illegible in the presence of large sets of taxa (**Figure 9*a***), so important differences in lineage must be manually highlighted. In biome analysis, the main objective is to determine the composition of samples (**Figure 9*c***) and how they change and evolve over time. Stacked bar plots (**Figure 9*c***) allow broad differences in composition to be shown, but it is important to also show lineage. This is often done with sunburst plots (**Figure 9*d***) – although using flame graphs instead makes it easier to compare multiple plots (**Figure 9*e***). Similarly, the Venn diagrams (**Figure 9*f***) typically used to visualize species co-occurrence can often be replaced by linear diagrams (**Figure 9*g***). Overall, there remains considerable scope for improvements – for example, with many of the current tools in this research area, considerable effort is required when using them to create figures for publication that are uncluttered, effective, and visually expressive.

Emerging frontiers. Advances in genomic sequencing allow us to examine differences within populations, and across ecosystems, in unprecedented detail. A plethora of microbiome sequence data promises to revolutionize our understanding of evolution and human health. But we are struggling to develop effective visual analysis strategies because no single visual metaphor captures the richness of population level data. Pan-genome visualizations, designed to show core and 'accessory' genes in a species' genome, are dashboards that combine existing methods (92). Phylogeography uses maps to show phylogenetic relationships across geocoded samples (93).

Pathogen surveillance, a crucial challenge which spans these fields, requires integrated, alternative views that capture population dynamics and highlight emerging resistance (94).

Our ability to observe biology at the molecular, cellular, anatomical and physiological level has never been greater; but making sense of these emerging data will require overcoming formidable analytical challenges, as well as the invention of fundamentally new, visual metaphors (95), changing how we see and think about our data.

VISUALIZATION FOR COMMUNICATION

Science is not complete until it is communicated (**Figure 1*a***); however, this is often challenging, due to the inherent complexity of biomedical research. Fortunately, visualization can help here as well.

Figures & Illustrations

In preparing a publication, visualization tools used for discovery are typically used to select static views that best express the insights found. A very small number of journals allow interactive figures; unfortunately, interactive figures in publications are often complex to produce and difficult to maintain – and, ironically, are used by very few readers. Interactive figures can augment static figures, but not replace them: just as scientific writing commits us to a particular way of describing our work, a static view commits us to a particular viewpoint that we believe best expresses the phenomena revealed by our data. Although difficult, selecting static views is often an essential step in research – and can lead to new insights.

In many cases, these static views need to be post-processed, using tools such as Illustrator or Photoshop, to improve clarity and ensure that marks and labels are consistent and readable at publication scale. In addition, since ~5% of readers and reviewers are colorblind (97), it is good practice to use color blindness proofing tools[5] and, where needed, modify figures to avoid relying on red-green contrast – this can usually be achieved by adjusting saturation and lightness values to increase contrast (97).

However, it sometimes can be unclear where the boundary lies between necessary improvements versus scientific fraud; thus, it is important to follow established guidelines on image and figure manipulation (96).

Animations & Videos

Animations and videos can dramatically enhance scientific communication and are becoming easier and cheaper to produce, leading to a marked increase in scientific video content (98). Done well, scientific videos improve peer-to-peer communication, and inspire public engagement and enthusiasm (27). Unfortunately, ensuring scientific accuracy typically involves considerable time and effort, as does achieving a high standard in video production, which requires learning cinematography principles, practices, and tools (e.g., Autodesk Maya, Blender, and Adobe After-Effects).

[5] In Illustrator & Photoshop, choose the menu items View > Proof Setup > Color Blindness to preview how a figure will appear to people with common forms of color blindness.

To help overcome these barriers, several specialist tools are being developed to streamline and simplify the production of scientific animations (99). Unfortunately, until these efforts become more advanced, accurate and compelling scientific videos will likely remain relatively rare.

PERSPECTIVES

This review has highlighted a few specific cases where data visualization is being used to accelerate discovery; but biomedical science has thousands more. Thus, while many visualization tools are already available (1, 21-25, 30), they are often inadequate for cutting-edge datasets. Addressing this challenge requires the invention of novel, tailored visualization strategies, each adapted to specific scenarios; this can be very difficult and, in many cases, is a rate limiting step in discovery. The resources outlined in this review can help (particularly **Table 1**). Especially noteworthy resources include the 2010 Nature Methods special issue on 'Visualizing Biological Data' (1, 21-25), and the ongoing Nature Methods 'Points of View' article series focused on specific visualization issues for life scientists (6). It can also be useful to exchange experiences with peers facing similar challenges: the annual Visualizing Biological Data (VIZBI) conference provides a forum for this exchange, and also provides a free, online collection of videos and posters from previous meetings (http://vizbi.org/). The VIZBI forum is also designed to help bioinformaticians connect with graphic designers, graphic artists, and biomedical communicators using illustration or animation. It can also help to connect with computer scientists researching data visualization: as well as advising on good principles and practice, they can be valuable collaborators[7]; a forum for such engagement is provided by the annual BioVis symposium (http://biovis.net/).

Tailored visualization tools play a critical role in research, some becoming widely used and highly cited (100). However, limitations in popular tools (e.g., cluttered, overly complicated user interfaces or poorly chosen defaults, such as rainbow color maps (11)) can have very negative impacts, contributing to dead-end research and incorrect diagnoses. Creating tailored tools with good visualization and design practices (101) typically requires years of sustained focus: it is often not clear how tool development and maintenance can be funded – however it is clear that this needs to be a central issue in research funding policies.

The survey in this review has highlighted that specific research areas need highly tailored visualization tools; nonetheless, there are common generic methods (e.g., tree graphs, parallel coordinates, stacked bar charts) and strategies (e.g., clustering, alternative views, overview/details, linked views, minimal coloring) being used across many research areas. There are also common outstanding challenges, such as uncertainty visualization (102) and multiscale navigation. Unfortunately, some poor visualization practices are also common (e.g., overly cluttered visualizations). However, there are good indicators that this situation is improving, and that there is increased awareness of the importance of data visualization in the life sciences (103); this is evidenced by the increased focus on visualization in mainstream conferences, as well as the emergence of more specialist meetings, such as VIZBI and BioVis.

As we seek to improve our tailored visualizations, another common challenge is how to objectively assess the quality of a particular visualization method or tool (104). An obvious and important, quality measurement is rate of adoption by the community; however, the popularity of a visualization strategy often has more to do with the cognitive load required to first learn how to read

[7] Computer scientists sometimes describe themselves as 'solution rich, and problem poor' – while biomedical researchers are generally 'problem rich, and solution poor'.

it, rather than how effectively or expressively it allows data to be understood and new insights to be generated. Here again, data visualization research can help: methods are being developed for quantitatively evaluating the effectiveness of visualizations (28) and for assessing visual information processing – for example, via eye-tracking (105) or via brain activity measurements to assess cognitive load (106). Hopefully, these evaluation methods will soon provide objective measures of the value of a visualization (104) that are recognized and agreed upon by the research community. This, together with other advances in data visualization and user experience design (101), may soon provide a new generation of tools that are much more powerful yet also easier to learn and use. Such tools would significantly reduce many of the current frustrations of scientists and clinicians, and revolutionize how we see and think about our data.

Conclusions: To understand and gain insight from the large, complex datasets generated in biomedical research, we need tailored visualization methods and tools that present the right data and analysis to the right researcher or clinician at the right time – providing a clear view of the inherent complexity in our data, not the complexity of oversimplification[9]. The development and adoption of such methods and tools will require fundamental changes to current research, communication, training, and clinical practices. Without these changes, many biomedical insights will remain undiscovered and misdiagnoses will remain unrecognized, buried in data already collected.

DISCLOSURE STATEMENT

None.

ACKNOWLEDGEMENTS

Much of this review emerged from discussions at VIZBI 2017 in Sydney, Australia. We are grateful for the many colleagues who have contributed, helping articulate the problems of specific areas in the biosciences and also to Bang Wong, Tamara Munzner, and Jeffery Heer, who have developed groundbreaking best practices. We are grateful for funding organizations that have supported the VIZBI meeting series, especially NIH, EMBO, the University of Sydney, and the Garvan Institute of Medical Research. We would also like to acknowledge useful discussion and comments with Drew Berry, Christian Stolte, David James, and Sean Humphrey.

SIDEBARS

Data Visualization, SciVis, and InfoVis

Computer scientists have long used the term ‘scientific visualization’ (or SciVis) to describe visualization of data that directly map into two or three spatial dimensions (e.g., cartography or computed tomography scans). In contrast, ‘information visualization’ (or InfoVis) is used to describe visualization of abstract data (e.g., classic 2D data plots, or network graphs). Since around the year 2000, ‘data visualization’ has emerged as a unifying term that encompasses both of these historically separated research fields.

[9] Paraphrased from Edward Tufte.

Avoiding Common Visualization Mistakes

Biomedical data is often difficult and expensive to acquire and analyze. Ironically, at the final step, when data are visualized, we often use visual techniques that obfuscate true patterns in our data and introduce visual artifacts. To avoid the most common mistakes:

- Avoid rainbow color maps (11, 107).
- Use color minimally (108). Color used poorly is worse than no color at all (109).
- Avoid creating confusing, overcrowded visualizations (e.g., 'hairball' graphs). Reduce information via filtering or clustering; or use a different layout (**Figure 7**).
- Avoid 3D visualizations for abstract data – use only for spatial data that is intrinsically 3D, e.g., macromolecular structures (110).
- Avoid conflating research and art. Many of the commonly-used tailored tools provide powerful features that make it easy to create visualizations that are dramatic or aesthetically appealing, but where the underlying scientifically meaning becomes obscured. This can be useful when creating impactful artwork (e.g., a cover figure); but it undermines the goals of data visualization in research, which are always clarity and insight.

Optical Illusions Caused by Ground Subtraction

Visualizations that rely on color to encode quantitative values are subject to an optical illusion known as 'ground subtraction'. In a heat map, for example, strongly contrasting colors in a cell's neighbors can make it appear much higher or lower in luminance than it should. This illusion can be very strong; as demonstrated in the 'checker shadow' image by Edward H. Adelson (http://persci.mit.edu/gallery/checkershadow), the human visual system can be surprisingly inaccurate at reading values encoded with color. In heat maps, this illusion can mask true patterns and introduce visual artifacts (46). As the number of rows and columns increase, or as cell size is reduced, the effect can become worse, making it impractical to display all significant results as one very large heat map.

Thus, for the display of quantities where absolute variation between observations is important, it is recommended to encode values with position or size rather than lightness, saturation, or color hue (**Figure 2**).

An Exemplary Biomedical Databank

Compared with many areas of biomedical science, visualization methods for macromolecular structures are more advanced – largely because they build upon a solid bedrock of exceptionally well-managed data. Created in 1972, the Protein Data Bank (PDB, 50) has exemplary practices and stability, as outlined below, that facilitate reproducibility and substantially simplify the difficult task of creating and maintaining tailored visualization tools. Unfortunately, most biomedical databanks created since have not learned from these practices, thus requiring tool developers to contend with data formats and sources that are many, varied, and often unstable.

- Each entry is a deposition related to one specific scientific publication, not to an abstract concept (e.g., a 'gene' or 'pathway') whose definition may change over time.

- Entry identifiers are short, and designed to be easy to remember. Depositors can propose an identifier, ensuring many are meaningful. For example, the first crystal structure of the protein ‘actin’ has identifier ‘1ATN’ (111).
- Entries include rich meta-data describing how they were generated (and by whom), and cross-linking to related databases.
- Raw and processed data are stored, enabling later re-analyses.
- An international network of organizations maintains and curates the database.
- PDB deposition is required when publishing in major journals.

TERMS AND DEFINITIONS

1. **Data visualization:** use of computer-aided, interactive, visual representations of data to amplify cognition (10) and accelerate discovery and communication.
2. **Data density:** the total number of data entries shown in a visualization divided by the display area.
3. **Visual effectiveness:** the accuracy and clarity with which a given visual encoding of data is conveyed to the reader.
4. **Multivariate data:** data comprised of multiple variables of any type, including quantitative, categorical (‘A is cytoplasmic’), or relational (‘A binds B’).
5. **Visual channel:** an elementary graphical strategy used for visually encoding data (e.g., using color hue to show data categories).
6. **Visual expressiveness:** how well a visualization expresses all – and only - information most relevant to the phenomena studied.
7. **Tailored visualization:** a strategy designed for integrating specific types of datasets, with supporting context, in a manner understood by peers.
8. **Spatial reasoning:** use of visual perception to enhance cognition; aided by organizing relevant data on a graphical display.
9. **Parallel coordinate plot:** a profile plot of multidimensional points, each shown as a series of line segments connecting parallel axes.
10. **Overview/details:** ‘overview first, zoom & filter, details on demand’ is the ‘Visual Information-Seeking Mantra’ for large datasets (112).
11. **Alternative views:** different ways of visualizing the same multidimensional dataset, each of which can provide different insights.
12. **Brushing and linking:** linked alternative views, where interactive changes made in one are automatically reflected in the other (113).
13. **Heat map:** a graphical representation of a matrix of data where individual values are encoded using color.
14. **t-distributed stochastic neighbor embedding (t-SNE):** a dimensionality reduction method aimed at preserving inter-point similarities and differences.

15. **Stacked bar chart:** a visualization in which bars representing related data are stacked on top of (or beside) each other.

16. **Virtual reality (VR):** blocking a person's view of their surroundings via head-mounted displays (e.g., Oculus), allowing immersion in artificially-generated content.

17. **Augmented reality (AR):** augmenting a person's normal view of their surroundings by adding computer-generated images or data (e.g., HoloLens).

18. **Tree graph:** a graph where all lines connect without forming loops; used for hierarchical data.

19. **Flame graph**: a visualization of hierarchical data where width encodes branch quantity, and sub-branches are stacked on parent branches.

20. **Linear diagram:** shows the size of overlaps and differences amongst multiple sets of data; an alternative to Venn diagrams.

LITERATURE CITED

1. O'Donoghue SI, Gavin A-C, Gehlenborg N, Goodsell DS, Hériché J-K, et al. 2010. Visualizing biological data - now and in the future. *Nature Methods* 7: S2-S4
2. Graber ML, Franklin N, Gordon R. 2005. Diagnostic error in internal medicine. *Arch Intern Med* 165: 1493-9
3. Pinto A, Brunese L. 2010. Spectrum of diagnostic errors in radiology. *World J Radiol* 2: 377-83
4. Makary MA, Daniel M. 2016. Medical error-the third leading cause of death in the US. *BMJ* 353: i2139
5. Tufte ER. 2009. *The Visual Display of Quantitative Information*. Cheshire: Graphics Press
6. Evanko D. 2013. Data visualization: A view of every points of view column. Methagora: a blog from Nature Methods. http://blogs.nature.com/methagora/2013/07/data-visualization-points-of-view.html
7. Rougier NP, Droettboom M, Bourne PE. 2014. Ten simple rules for better figures. *PLoS Comput Biol* 10: e1003833
8. Munzner T. 2014. *Visualization Analysis and Design*: CRC Press
9. Ware C. 2004. *Information visualization: Perception for design*. San Francisco, USA: Morgan-Kauffman
10. Card SK, Mackinlay JD, Shneiderman B. 1999. *Readings in information visualization: using vision to think*: Morgan Kaufmann
11. Borland D, Taylor MR, 2nd. 2007. Rainbow color map (still) considered harmful. *IEEE Comput Graph Appl* 27: 14-7
12. Craft M, Dobrenz B, Dornbush E, Hunter M, Morris J, et al. 2015. *An assessment of visualization tools for patient monitoring and medical decision making*. Presented at Systems and Information Engineering Design Symposium (SIEDS), 2015
13. Lewandowsky S, Spence I. 1989. The perception of statistical graphs. *Sociological Methods & Research* 18: 200-42
14. Koch K, McLean J, Segev R, Freed MA, Berry MJ, 2nd, et al. 2006. How much the eye tells the brain. *Curr Biol* 16: 1428-34
15. Healey CG, Enns JT. 2012. Attention and visual memory in visualization and computer graphics. *IEEE Trans Vis Comput Graph* 18: 1170-88
16. Ball R, North C. 2007. Realizing embodied interaction for visual analytics through large displays. *Computers & Graphics* 31: 380-400
17. Cleveland WS, Diaconis P, McGill R. 1982. Variables on Scatterplots Look More Highly Correlated When the Scales are Increased. *Science* 216: 1138-41
18. Heer J, Kong N, Agrawala M. 2009. *Sizing the horizon: the effects of chart size and layering on the graphical perception of time series visualizations*. Presented at Proceedings of the SIGCHI Conference on Human Factors in Computing Systems
19. Inselberg A. 1997. *Multidimensional detective*. Presented at Information Visualization, 1997. Proceedings., IEEE Symposium on
20. Hegarty M. 2011. The cognitive science of visual-spatial displays: Implications for design. *Topics in cognitive science* 3: 446-74
21. Nielsen CB, Cantor M, Dubchak, I., Gordon D, Wang T. 2010. Visualizing Genomes: Techniques and Challenges. *Nature Methods* 7: S5-S15
22. Procter JB, Barton GJ, Thompson J, Westhof E, Creevey C, Letunic I. 2010. Visualization of multiple alignments, phylogenies and gene family evolution. *Nature Methods* 7: S16-S25
23. O'Donoghue SI, Goodsell DS, Frangakis AS, Jossinet F, Laskowski R, et al. 2010. Visualization of macromolecular structures. *Nature Methods* 7: S42-S55
24. Gehlenborg N, O'Donoghue SI, Baliga NS, Goesmann A, Hibbs MA, et al. 2010. Visualization of omics data for systems biology. *Nature Methods* 7: S56-S68
25. Walter T, Shattuck D, Baldock R, Bastin M, Carpenter AE, et al. 2010. Visualization of image data from cells to organisms. *Nature Methods* 7: S26-S41

26. Soegaard M, Rikke Friis D, eds. 2013. *The Encyclopedia of Human-Computer Interaction*. Aarhus, Denmark: The Interaction Design Foundation. https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed
27. Johnson GT, Hertig S. 2014. A guide to the visual analysis and communication of biomolecular structural data. *Nat Rev Mol Cell Biol* 15: 690-8
28. Heer J, Bostock M. 2010. Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design. In *CHI*. Atlanta, Georgia, USA
29. Inselberg A. 2009. *Parallel coordinates: Visual multidimensional geometry and its applications*. New York, USA: Springer
30. Pavlopoulos GA, Malliarakis D, Papanikolaou N, Theodosiou T, Enright AJ, Iliopoulos I. 2015. Visualizing genome and systems biology: technologies, tools, implementation techniques and trends, past, present and future. *GigaScience* 4: 38
31. Ernst J, Kellis M. 2012. ChromHMM: automating chromatin-state discovery and characterization. *Nature methods* 9: 215-16
32. Pabinger S, Dander A, Fischer M, Snajder R, Sperk M, et al. 2014. A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in bioinformatics* 15: 256-78
33. Schroeder MP, Gonzalez-Perez A, Lopez-Bigas N. 2013. Visualizing multidimensional cancer genomics data. *Genome medicine* 5: 9
34. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326: 289-93
35. Serra F, Di Stefano M, Spill YG, Cuartero Y, Goodstadt M, et al. 2015. Restraint-based three-dimensional modeling of genomes and genomic domains. *FEBS Lett* 589: 2987-95
36. Ay F, Noble WS. 2015. Analysis methods for studying the 3D architecture of the genome. *Genome Biol* 16: 183
37. Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, et al. 2016. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell systems* 3: 99-101
38. Zhou X, Li D, Lowdon RF, Costello JF, Wang T. 2014. methylC Track: visual integration of single-base resolution DNA methylation data on the WashU EpiGenome Browser. *Bioinformatics* 30: 2206-07
39. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, et al. 2009. Circos: an information aesthetic for comparative genomics. *Genome Res* 19: 1639-45
40. Taberlay PC, Achinger-Kawecka J, Lun ATL, Buske FA, Sabir KS, et al. 2016. Three-dimensional disorganisation of the cancer genome occurs coincident with long range genetic and epigenetic alterations. *Genome Research* 26: 719-31
41. Shalon D, Smith SJ, Brown PO. 1996. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res* 6: 639-45
42. Wills QF, Livak KJ, Tipping AJ, Enver T, Goldson AJ, et al. 2013. Single-cell gene expression analysis reveals genetic associations masked in whole-tissue experiments. *Nat Biotechnol* 31: 748-52
43. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, et al. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320: 1344-9
44. Gierlinski M, Cole C, Schofield P, Schurch NJ, Sherstnev A, et al. 2015. Statistical models for RNA-seq data derived from a two-condition 48-replicate experiment. *Bioinformatics* 31: 3625-30
45. Wilkinson L, Friendly M. 2009. The history of the cluster heat map. *The American Statistician* 63: 179-84
46. Wong B. 2010. Points of view: Color coding. *Nature Methods* 7: 573-73
47. Pereverzeva M, Murray SO. 2014. Luminance gradient configuration determines perceived lightness in a simple geometric illusion. *Front Hum Neurosci* 8: 977

48. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, et al. 2014. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* 32: 381-86

49. Cao J, Packer JS, Ramani V, Cusanovich DA, Huynh C, et al. 2017. Comprehensive single cell transcriptional profiling of a multicellular organism by combinatorial indexing. *bioRxiv*

50. Berman H, Henrick K, Nakamura H. 2003. Announcing the worldwide Protein Data Bank. *Nat Struct Biol* 10: 980

51. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, et al. 2004. UCSF Chimera - a visualization system for exploratory research and analysis. *J Comput Chem* 25: 1605-12

52. Kozlikova B, Krone M, Lindow N, Falk M, Baaden M, et al. 2015. *Visualization of biomolecular structures: State of the art*. Presented at Eurographics Conference on Visualization (EuroVis)-STARs

53. Kendrew J, Dickerson R, Strandberg B, Hart R, Davies D, et al. 1960. Structure of myoglobin: A three-dimensional Fourier synthesis at 2 Å. resolution. *Nature* 185: 422-27

54. Farrugia L. 2012. WinGX and ORTEP for Windows: an update. *Journal of Applied Crystallography* 45: 849-54

55. Chung JC, Harris MR, Brooks FP, Fuchs H, Kelley MT, et al. 1989. *Exploring virtual worlds with head-mounted displays*. Presented at Three-Dimensional Visualization and Display Technologies, Los Angeles, USA

56. Humphrey W, Dalke A, Schulten K. 1996. VMD: visual molecular dynamics. *J. Mol. Graph.* 14: 33-38

57. Gillet A, Sanner M, Stoffler D, Olson A. 2005. Tangible interfaces for structural molecular biology. *Structure* 13: 483-91

58. Sabir KS, Stolte C, Tabor B, O'Donoghue SI. 2013. *The Molecular Control Toolkit: Controlling 3D molecular graphics via gesture and voice*. Presented at IEEE Symposium on Biological Data Visualization (BioVis), Atlanta, GA, USA

59. Gillet A, Sanner M, Stoffler D, Goodsell D, Olson A. 2004. *Augmented reality with tangible auto-fabricated models for molecular biology applications*. Presented at Visualization, 2004. IEEE

60. Heinrich J, Vuong J, Hammang CJ, Wu A, Rittenbruch M, et al. 2016. Evaluating viewpoint entropy for ribbon representation of protein structure. *Computer Graphics Forum* 35: 181–90

61. Lv Z, Tek A, Da Silva F, Empereur-Mot C, Chavent M, Baaden M. 2013. Game on, science-how video game technology may help biologists tackle visualization challenges. *PloS one* 8: e57990

62. O'Donoghue SI, Sabir KS, Kalemanov M, Stolte C, Wellmann B, et al. 2015. Aquaria: Simplifying discovery and insight from protein structures. *Nature Methods* 12: 98–99

63. Levitt M. 2009. Nature of the protein universe. *Proceedings of the National Academy of Sciences of the United States of America* 106: 11079-84

64. Perdigão N, Heinrich J, Stolte C, Sabir KS, Buckley MJ, et al. 2015. Unexpected features of the dark proteome. *Proceedings of the National Academy of Sciences of the United States of America* 112: 15898–903

65. Rysavy SJ, Beck DA, Daggett V. 2014. Dynameomics: Data-driven methods and models for utilizing large-scale protein structure repositories for improving fragment-based loop prediction. *Protein Science* 23: 1584-95

66. Bai X-C, McMullan G, Scheres SH. 2015. How cryo-EM is revolutionizing structural biology. *Trends in biochemical sciences* 40: 49-57

67. Johnson GT, Autin L, Al-Alusi M, Goodsell DS, Sanner MF, Olson AJ. 2015. cellPACK: a virtual mesoscope to model and visualize structural systems biology. *Nature methods* 12: 85-91

68. Ghosh S, Matsuoka Y, Asai Y, Hsin K-Y, Kitano H. 2011. Software for systems biology: from tools to integrated platforms. *Nature reviews. Genetics* 12: 821

69. Kitano H. 2002. Systems biology: a brief overview. *Science* 295: 1662-64
70. Schomburg D, Michal G. 2012. *Biochemical pathways : an atlas of biochemistry and molecular biology*. Hoboken, N.J.: John Wiley & Sons. xi, 398 p. pp.
71. Bader GD, Cary MP, Sander C. 2006. Pathguide: a pathway resource list. *Nucleic acids research* 34: D504-D06
72. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, et al. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 13: 2498-504
73. Bastian M, Heymann S, Jacomy M. 2009. Gephi: an open source software for exploring and manipulating networks. *Icwsm* 8: 361-62
74. Kobourov SG. 2013. In *Handbook of Graph Drawing and Visualization*, ed. R Tamassia, pp. 383-408: CRC Press
75. Barsky A, Gardy JL, Hancock RE, Munzner T. 2007. Cerebral: a Cytoscape plugin for layout of and interaction with biological networks using subcellular localization annotation. *Bioinformatics* 23: 1040-2
76. Holten D, Van Wijk JJ. 2009. *Force-Directed Edge Bundling for Graph Visualization*. Presented at Computer graphics forum
77. Barsky A, Munzner T, Gardy J, Kincaid R. 2008. Cerebral: Visualizing multiple experimental conditions on a graph with biological context. *IEEE transactions on visualization and computer graphics* 14: 1253-60
78. Zhou H-X, Rivas G, Minton AP. 2008. Macromolecular crowding and confinement: biochemical, biophysical, and potential physiological consequences. *Annu. Rev. Biophys.* 37: 375-97
79. Von Landesberger T, Kuijper A, Schreck T, Kohlhammer J, van Wijk JJ, et al. 2011. *Visual analysis of large graphs: state-of-the-art and future research challenges*. Presented at Computer graphics forum
80. Kwon O-H, Crnovrsanin T, Ma K-L. 2017. What Would a Graph Look Like in This Layout? A Machine Learning Approach to Large Graph Visualization. *IEEE Transactions on Visualization and Computer Graphics*
81. Aebersold R, Mann M. 2016. Mass-spectrometric exploration of proteome structure and function. *Nature* 537: 347-55
82. Humphrey SJ, Azimifar SB, Mann M. 2015. High-throughput phosphoproteomics reveals in vivo insulin signaling dynamics. *Nat Biotechnol* 33: 990-5
83. Ma DK, Stolte C, Krycer JR, James DE, O'Donoghue SI. 2015. SnapShot: Insulin/IGF1 Signaling. *Cell* 161: 948-48.e1
84. Burgess A, Vuong J, Rogers S, Malumbres M, O'Donoghue SI. 2017. SnapShot: Phosphoregulation of mitosis. *Cell* 169: 1358–58.e1
85. Sydor AM, Czymmek KJ, Puchner EM, Mennella V. 2015. Super-Resolution Microscopy: From Single Molecules to Supramolecular Assemblies. *Trends Cell Biol* 25: 730-48
86. Reynaud EG, Peychl J, Huisken J, Tomancak P. 2015. Guide to light-sheet microscopy for adventurous biologists. *Nat Methods* 12: 30-4
87. Walter T, Shattuck DW, Baldock R, Bastin ME, Carpenter AE, et al. 2010. Visualization of image data from cells to organisms. *Nat Methods* 7: S26-41
88. Rossner M, Yamada KM. 2004. What's in a picture? The temptation of image manipulation. *J Cell Biol* 166: 11-5
89. Burel J-M, Besson S, Blackburn C, Carroll M, Ferguson RK, et al. 2015. Publishing and sharing multi-dimensional image data with OMERO. *Mammalian Genome* 26: 441-47
90. Bernhardt S, Nicolau SA, Soler L, Doignon C. 2017. The status of augmented reality in laparoscopic surgery as of 2016. *Med Image Anal* 37: 66-90
91. Maier-Hein L, Vedula SS, Speidel S, Navab N, Kikinis R, et al. 2017. Surgical data science for next-generation interventions. *Nature Biomedical Engineering* 1: 691
92. Ding W, Neher R. 2014. http://pangenome.tuebingen.mpg.de/. pp. PanX Pangenome visualisation tool

93. Parks DH, Porter M, Churcher S, Wang S, Blouin C, et al. 2009. GenGIS: A geospatial information system for genomic data. *Genome Res* 19: 1896-904
94. Argimon S, Abudahab K, Goater RJ, Fedosejev A, Bhai J, et al. 2016. Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. *Microb Genom* 2: e000093
95. Paten B, Novak AM, Eizenga JM, Garrison E. 2017. Genome graphs and the evolution of genome inference. *Genome Res* 27: 665-76
96. Cromey DW. 2010. Avoiding twisted pixels: ethical guidelines for the appropriate use and manipulation of scientific digital images. *Science and engineering ethics* 16: 639-67
97. Wong B. 2011. Color blindness. *Nat Methods* 8: 441
98. McGill G. 2008. Molecular movies... coming to a lecture near you. *Cell* 133: 1127-32
99. Iwasa JH. 2015. Bringing macromolecular machinery to life using 3D animation. *Current opinion in structural biology* 31: 84-88
100. Van Noorden R, Maher B, Nuzzo R. 2014. The top 100 papers. *Nature* 514: 550-3
101. Shneiderman B. 2010. *Designing the user interface: strategies for effective human-computer interaction*: Pearson Education India
102. Sanyal J, Zhang S, Bhattacharya G, Amburn P, Moorhead R. 2009. A user study to compare four uncertainty visualization methods for 1d and 2d datasets. *IEEE transactions on visualization and computer graphics* 15
103. Callaway E. 2016. The visualizations transforming biology. *Nature* 535: 187-8
104. Van Wijk JJ. 2005. *The value of visualization*. Presented at Visualization, 2005. VIS 05. IEEE
105. Blascheck T, Kurzhals K, Raschke M, Burch M, Weiskopf D, Ertl T. 2014. *State-of-the-art of visualization for eye tracking data*. Presented at Proceedings of EuroVis
106. Anderson EW, Potter KC, Matzen LE, Shepherd JF, Preston GA, Silva CT. 2011. *A user study of visualization effectiveness using EEG and cognitive load*. Presented at Computer Graphics Forum
107. Gehlenborg N, Wong B. 2012. Mapping quantitative data to color: data structure informs choice of color maps. *Nature Methods* 9: 769-70
108. Wong B. 2011. Points of view: avoiding color. *nature methods* 8: 525-25
109. Tufte ER. 1990. *Envisioning Information*. Chesire: Graphics Press
110. Gehlenborg N, Wong B. 2012. Points of view: Into the third dimension. *Nature methods* 9: 851-51
111. Kabsch W, Mannherz HG, Suck D, Pai EF, Holmes KC. 1990. Atomic structure of the actin:DNase I complex. *Nature* 347: 37-44
112. Shneiderman B. 1996. *The eyes have it: A task by data type taxonomy for information visualizations*. Presented at Visual Languages, 1996. Proceedings., IEEE Symposium on
113. Buja A, McDonald JA, Michalak J, Stuetzle W. 1991. *Interactive data visualization using focusing and linking*. Presented at Visualization, 1991. Visualization'91, Proceedings., IEEE Conference on
114. Wolfram Research I. 2017. *Mathematica*. Champaign, Illinois: Wolfram Research, Inc.
115. Hunter JD. 2007. Matplotlib: A 2D graphics environment. *Computing In Science & Engineering* 9: 90-95
116. Wickham H. 2016. *ggplot2: elegant graphics for data analysis*: Springer
117. Anscombe FJ. 1973. Graphs in Statistical Analysis. *American Statistician* 27: 17-21
118. Matejka J, Fitzmaurice G. 2017. Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing. In *CHI*. Denver, CO, USA: ACM
119. Fry B. 2008. *Visualizing Data*: O'Reilly Media
120. Moreland K. 2016. *Why We Use Bad Color Maps and What You Can Do About It*. Presented at IS&T International Symposium on Electronic Imaging, San Francisco, USA
121. Mackinlay JD. 1986. Automating the design of graphical presentations of relational information. *ACM Transactions on Graphics* 5

122. Gehlenborg N, Wong B. 2012. Points of view: heat maps. *Nature Methods* 9: 213-13

123. Yates A, Akanni W, Amode MR, Barrell D, Billis K, et al. 2015. Ensembl 2016. *Nucleic acids research* 44: D710-D16

124. Down TA, Piipari M, Hubbard TJ. 2011. Dalliance: interactive genome viewing on the web. *Bioinformatics* 27: 889-90

125. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, et al. 2014. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* 159: 1665-80

126. van der Maaten L, Hinton G. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9: 2579-605

127. Wattenberg M, Viégas F, Johnson I. 2016. How to Use t-SNE Effectively. *Distill* 1: e2

128. Haghverdi L, Buettner F, Theis FJ. 2015. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* 31: 2989-98

129. Coifman RR, Lafon S, Lee AB, Maggioni M, Nadler B, et al. 2005. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the National Academy of Sciences of the United States of America* 102: 7426-31

130. McCarthy DJ, Campbell KR, Lun AT, Wills QF. 2017. Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. *Bioinformatics* 33: 1179-86

131. Camp JG, Sekine K, Gerber T, Loeffler-Wirth H, Binder H, et al. 2017. Multilineage communication regulates human liver bud development from pluripotency. *Nature* 546: 533-38

132. Richardson JS, Richardson D, Tweedy N, Gernert K, Quinn T, et al. 1992. Looking at proteins: representations, folding, packing, and design. Biophysical Society National Lecture, 1992. *Biophysical journal* 63: 1185

133. Heinrich J, Kaur S, O'Donoghue SI. 2015. Evaluating the effectiveness of color to convey uncertainty in macromolecular structures. In *IEEE Symposium on Big Data Visual Analytics*, pp. 1-18. Hobart, Australia: IEEE

134. Dosztanyi Z, Csizmok V, Tompa P, Simon I. 2005. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics* 21: 3433-4

135. Rose AS, Bradley AR, Valasatava Y, Duarte JM, Prlić A, Rose PW. 2016. *Web-based molecular graphics for large complexes*. Presented at Proceedings of the 21st International Conference on Web3D Technology

136. Zhao G, Perilla JR, Yufenyuy EL, Meng X, Chen B, et al. 2013. Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics. *Nature* 497: 643-6

137. El Omari K, De Mesmaeker J, Karia D, Ginn H, Bhattacharya S, Mancini EJ. 2012. Structure of the DNA-bound T-box domain of human TBX1, a transcription factor associated with the DiGeorge syndrome. *Proteins: Structure, Function, and Bioinformatics* 80: 655-60

138. Isberg V, Mordalski S, Munk C, Rataj K, Harpsøe K, et al. 2015. GPCRdb: an information system for G protein-coupled receptors. *Nucleic acids research* 44: D356-D64

139. Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, et al. 2001. Intrinsically disordered protein. *Journal of Molecular Graphics and Modelling* 19: 26-59

140. Humphrey SJ, Yang G, Yang P, Fazakerley DJ, Stöckli J, et al. 2013. Dynamic adipocyte phosphoproteome reveals that Akt directly regulates mTORC2. *Cell Metab* 17: 1009-20

141. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, et al. 2015. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* 43: D447-52

142. Lu CP, Hager GD, Mjolsness E. 2000. Fast and globally convergent pose estimation from video images. *Ieee Transactions on Pattern Analysis and Machine Intelligence* 22: 610-22

143. Porter IM, McClelland SE, Khoudoli GA, Hunter CJ, Andersen JS, et al. 2007. Bod1, a novel kinetochore protein required for chromosome biorientation. *J Cell Biol* 179: 187-97

144. Nolden M, Zelzer S, Seitel A, Wald D, Müller M, et al. 2013. The Medical Imaging Interaction Toolkit: challenges and advances: 10 years of open-source development. *Int J Comput Assist Radiol Surg* 8: 607-20

145. Simpfendörfer T, Baumhauer M, Müller M, Gutt CN, Meinzer HP, et al. 2011. Augmented reality visualization during laparoscopic radical prostatectomy. *J Endourol* 25: 1841-5

146. Han MV, Zmasek CM. 2009. phyloXML: XML for evolutionary biology and comparative genomics. *Bmc Bioinformatics* 10

147. Darling ACE, Mau B, Blattner FR, Perna NT. 2004. Mauve: Multiple alignment of conserved genomic sequence with rearrangements. *Genome Research* 14: 1394-403

148. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, et al. 2010. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* 7: 335-36

149. Rodgers P, Stapleton G, Chapman P. 2015. Visualizing Sets with Linear Diagrams. *ACM Transactions on Computer-Human Interaction* 22: 27:1-27:39

150. Caporaso JG, Lauber CL, Costello EK, Berg-Lyons D, Gonzalez A, et al. 2011. Moving pictures of the human microbiome. *Genome Biology* 12

ANNOTED REFERENCES

1. (1) Nature Methods special issue on Visualizing Biological Data, covering molecular biology, biomedical science, and evolution.

2. (5). Inspirational, ground-breaking collection of historical and modern approaches to displaying quantitative data.

3. (6). Nature Methods regularly publishes 1-page articles focused on specific visualization issues faced by life scientists.

4. (7). Concise, practical guide to principles and tools for creating scientific figures.

5. (8). Comprehensive overview of data visualization principles.

6. (9). Outline of key principles and methods for interactive display of visual information.

7. (10) Definitive, annotated guide to classic papers on information visualization.

8. (27). Visual analysis and communication guide for biomolecular data – also relevant to other biomedical data.

9. (28) Groundbreaking method using Amazon's Mechanical Turk crowdsourcing platform to evaluate the effectiveness of visual encoding.

10. (29). Definitive guide to the theory and practice of using parallel coordinates to explore high-dimensional data.

TABLES

Table 1 Data visualization resources recommended for biomedical scientists in any field[10]

Resource	Description	URL
Discovery[11]		
Excel[$]	Everyday tool for generic visualization of smaller datasets	http://microsoft.com/excel
plotly	Online tool for fast data visualization	https://plot.ly/create/
Tableau[$]	For interactive visualizations, including web-based	http://tableau.com
Spotfire[$]	For visual analysis of larger datasets and tool generation	https://spotfire.tibco.com/
Origin[$,W]	For visual analysis of larger datasets	http://originlab.com
Mathematica[$]	For visual analysis of datasets & mathematical functions (114)	http://wolfram.com
MATLAB[$]	For visual analysis of datasets & mathematical functions	http://mathworks.com
Matplotlib	For tailored visualizations of datasets in Python (115)	http://matplotlib.org
ggplot2	For tailored visualizations of large, complex, datasets in R (116)	http://ggplot2.org
D3js	For tailored, interactive, web visualizations	http:// bit.ly/D3gallery
Communication		
Photoshop[$]	For editing imaging data	http://adobe.com/products
GIMP	Free, open source alternative to Photoshop	http://www.gimp.org
Illustrator[$]	For creating & editing vector graphics	http://adobe.com/Illustrator
Inkscape	Free, open source alternative to Illustrator	http://inkscape.org
MolecularMaya	Molecular structure plugin for Autodesk Maya[$] animation suite	http://bit.ly/molmaya
BioBlender	Molecular structure plugin for Blender animation suit	http://bioblender.org
Utilities		
Color Brewer	Web tool for selecting contrasting color maps	http://colorbrewer2.org
Adobe Color	Web tool for designing sets of colors	http://color.adobe.com
Paletton	Web tool for designing sets of colors	http://paletton.com
General Resources		
BioVis	Computer science publications on biological visualizations	http://biovis.net
Clarafi[$]	Training guides for biomedical visualization tools	http://clarafi.com
Info. is Beaut.	Showcase of charts and infographics for a wide variety of data	http://bit.ly/Info_Beauty
Vis. Complex.	Catalogue of tailored visualizations for complex data	http://visualcomplexity.com
VIZBI	Collected videos & posters on tailored biological visualizations	http://vizbi.org
Exemplars		
PDB101	Outstanding visual explanations of protein function and structure	https://pdb101.rcsb.org
Roche pathway	Tailored visualization showing ~3,000 metabolic reactions (70)	http://bit.ly/RochePathway
WEHI.tv	Collection of inspiring, informative biomedical animations	http://wehi.tv

[10] This table covers only tools and online resources; published articles and books describing generally useful visualization methods are highlighted as annotated references in **LITERATURE CITED**. Visualization methods tailored for specific fields of biomedical research are given in the corresponding sub-sections in **VISUALIZATION FOR DISCOVERY**.

[11] Arranged in approximate order starting at the top with ease-to-use, everyday tools for generic, once-off visualization tasks, and progressing to tools requiring more time and effort to use, but can manage large, complex data or reoccurring tasks.

[$] These tools cost money to use; the rest are free.

[W] Requires Microsoft Windows.

FIGURES

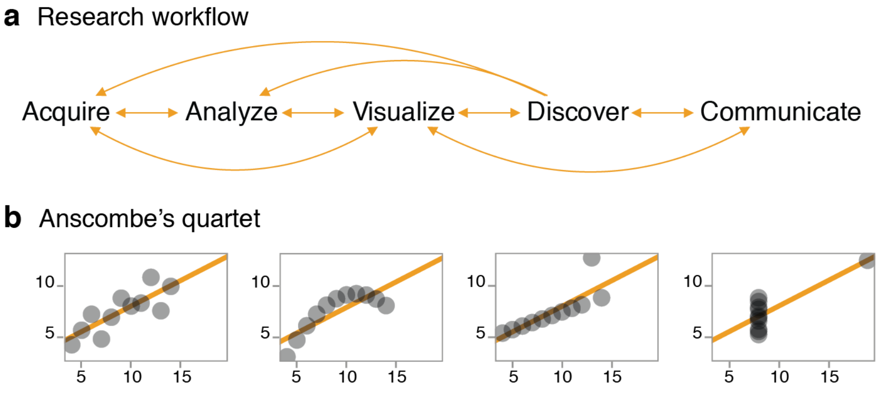


Figure 1

Role of data visualization in research. (*a*) As shown in this simplified model of the research workflow, data visualization is often a necessary and rate-limiting step in both discovery and communication. (*b*) Anscombe's quartet (117, 118) is a set of four 2D datasets in which *X* and *Y* values have identical mean, variance, and correlation coefficient. They also fit an identical linear regression line (orange) with identical coefficient of determination. Based on these statistics alone, we might expect all plots to be similar to the first; but visualization reveals surprisingly distinct patterns in each dataset. This demonstrates that we cannot skip from analysis to discovery: it is almost always necessary to confirm insights from automated analysis by manually visualizing data. Image *a* was adapted from Ben Fry (119).

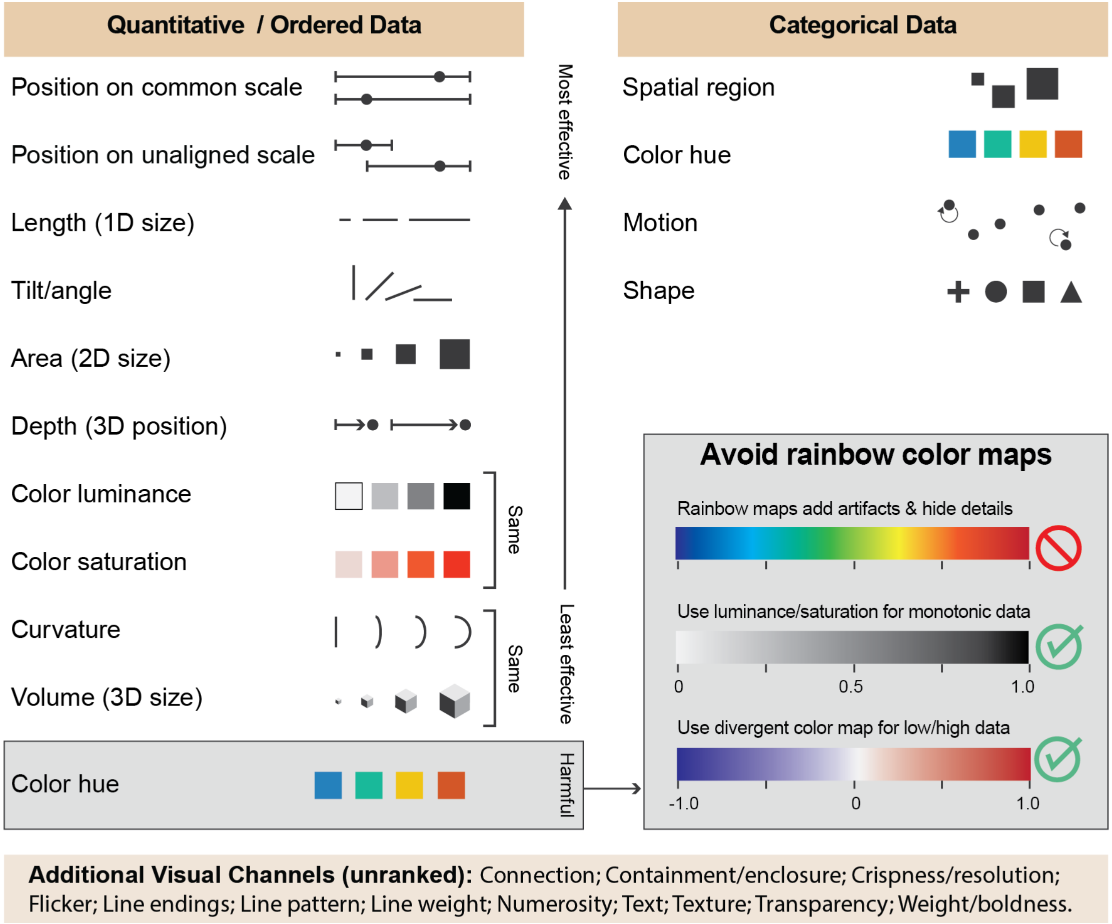
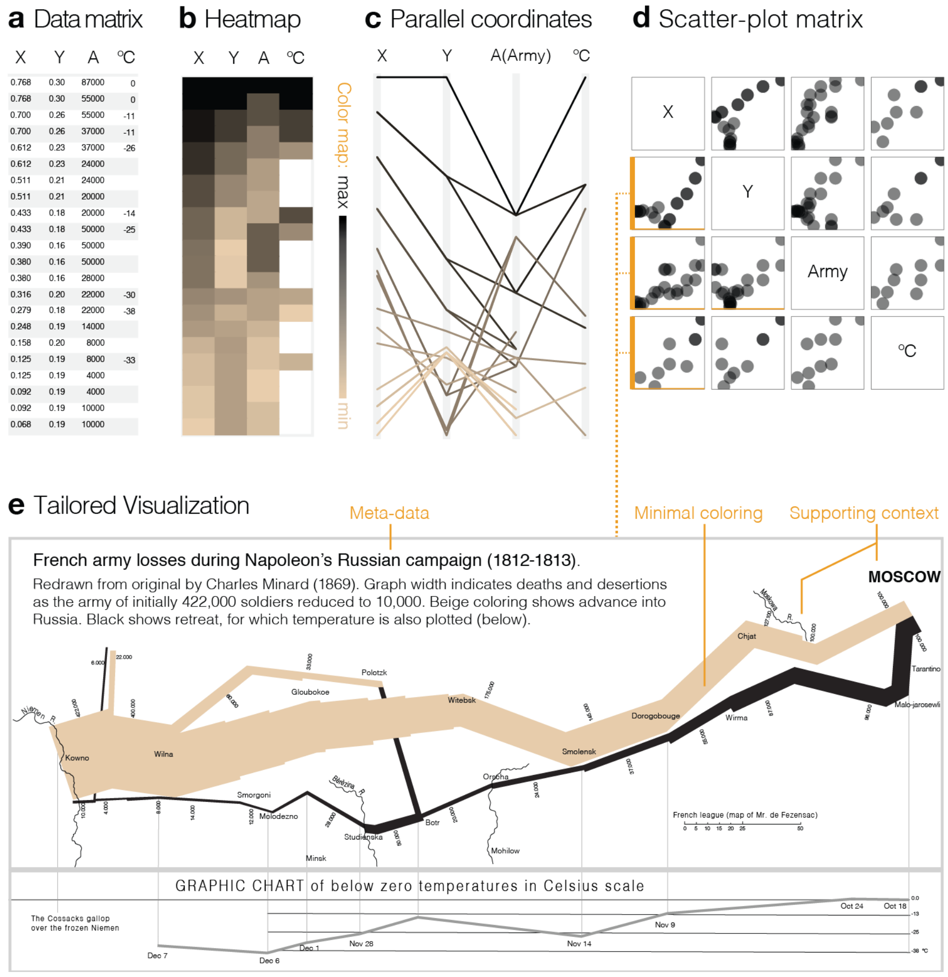

Quantitative / Ordered Data
Position on common scale
Position on unaligned scale
Length (1D size)
Tilt/angle
Area (2D size)
Depth (3D position)
Color luminance
Color saturation
Curvature
Volume (3D size)
Same
Same
Most effective
Least effective
Color hue
Harmful
Categorical Data
Spatial region
Color hue
Motion
Shape
Avoid rainbow color maps
Rainbow maps add artifacts & hide details
Use luminance/saturation for monotonic data
0
0.5
1.0
Use divergent color map for low/high data
-1.0
0
1.0
Additional Visual Channels (unranked): Connection; Containment/enclosure; Crispness/resolution; Flicker; Line endings; Line pattern; Line weight; Numerosity; Text; Texture; Transparency; Weight/boldness.

Figure 2

Visual channel ranking. Shows the most to least effective visual channels (top to bottom, respectively) for encoding quantitative, ordered, or categorical data. These rankings can be a useful guide in designing new visualizations and in studying exemplary visualizations (e.g., **Figure 3 & Figure 7*a***). Unfortunately, the use of color hue to encode quantitative data is still widespread in science, even though visualization research clearly demonstrates that this is not just ineffective, but can be harmful, introducing visual artifacts and hiding details (11). Instead (bottom right insert), use more effective color maps tailored for specific data ranges (120). Image was adapted from Munzner (8), based on the approach pioneered by Mackinlay (121) and extended by others (28).



X	Y	A	°C
0.768	0.30	87000	0
0.768	0.30	55000	0
0.700	0.26	55000	-11
0.700	0.26	37000	-11
0.612	0.23	37000	-26
0.612	0.23	24000	
0.511	0.21	24000	
0.511	0.21	20000	
0.433	0.18	20000	-14
0.433	0.18	50000	-25
0.390	0.16	50000	
0.380	0.16	50000	
0.380	0.16	28000	
0.316	0.20	22000	-30
0.279	0.18	22000	-38
0.248	0.19	14000	
0.158	0.20	8000	
0.125	0.19	8000	-33
0.125	0.19	4000	
0.092	0.19	4000	
0.092	0.19	10000	
0.068	0.19	10000	

French army losses during Napoleon's Russian campaign (1812-1813).

Redrawn from original by Charles Minard (1869). Graph width indicates deaths and desertions as the army of initially 422,000 soldiers reduced to 10,000. Beige coloring shows advance into Russia. Black shows retreat, for which temperature is also plotted (below).

Figure 3

Different 2D views of a 4D dataset. (*a*) Showing a data matrix as text reveals all information in 2D, but patterns can be difficult to detect. The data here cover the return journey shown in *e*, below (black). (*b*) In a heat map (122), the data matrix is visually encoded using color. This can give high data density; however, revealing data patterns often requires reordering rows and columns. Unfortunately, optical illusions can mask true patterns and introduce visual artifacts (46). (*c*) In parallel coordinate plots (19, 29), columns of the data matrix are represented as parallel axes, and each row becomes a series of line segments joining the axes. Correlations between adjunct axes can be easy seen – but not between non-adjacent axes. Hence, revealing data patterns often requires reordering axes, and careful choice of a color map (here, used to show X values). Parallel coordinates can reveal patterns not easily seen in a heat map – but they are usually not as compact. (*d*) A scatter-plot matrix shows every pairwise combination of columns in the data matrix, thus visualizing all two-dimensional correlations. (*e*) An exemplary tailored visualization of multivariate data, demonstrating many best practices. Note the use of very effective visual encodings (Figure 2). Note also what has been left out: color is used minimally, to distinguish the advance and retreat of the army (perhaps the most single most poignant categorical variable in this tragic data story); also, only the most visually expressive parts of the scatter-plot matrix in *d* (dotted line) have been included. Supporting context has been added via geographic features. Meta-data establishes credibility by identifying author, evidence sources, and methods. Tailored visualizations sometimes require using less effective channels: e.g., here, army size would be more effectively encoded using a bar chart – but this would not be as visually expressive. Similarly, it can sometimes be required to break visual conventions: e.g., here, the bottom plot implicitly shows time flowing right to left, with irregularly spaced dates. In spite of these departures from recommended guidelines, this tailored visualization "… may well be the best statistical graph ever drawn" (5). Image in *b* was made using

Excel, *c* & *d* using Matplotlib (115), and *e* was redrawn from Charles Minard's original using Illustrator.

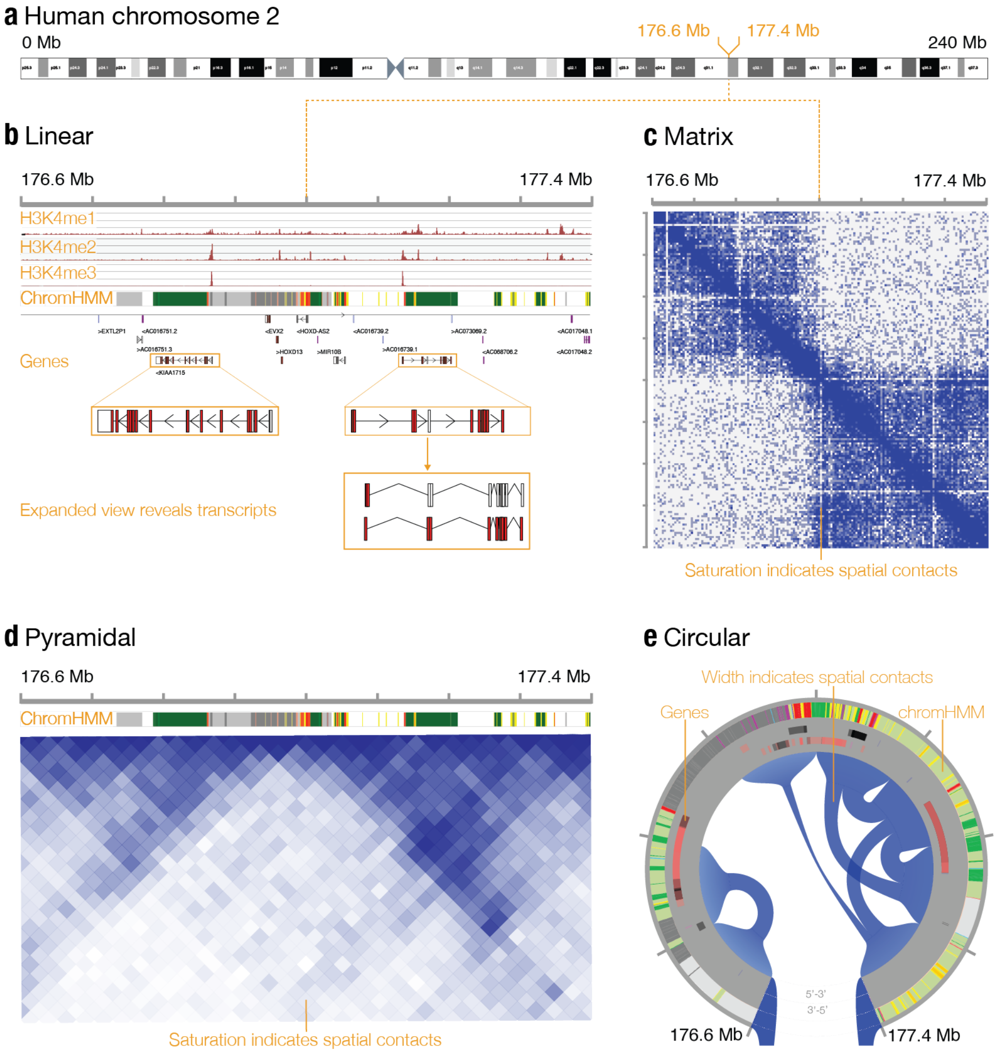


Figure 4

Genomic features of human chromosome 2. (*a*) The linear organization of chromosomes provides a natural visual layout for mapping genomic features such as cytobands, shown in this overview of a 240 megabase (Mb) chromosome. (*b*) Genome browsers enable navigation to small, specific regions that often contain vast numbers of features, including epigenetic marks (H3K4me1, etc.), genes, and regulatory elements. Graphical overviews for features are created with clustering methods such as ChromHMM (31), which condense many features into a single track, using color to indicate regions with similar features. Genes are also used as a graphical overview for the often-large number of transcripts they encode, which can be revealed upon demand. (*c*) Some features do not fit a linear layout; for example, Hi-C data (34), shown here, indicate 3D spatial contacts between genomic regions and can be encoded with color saturation and a contact matrix layout. (*d*) Rotating the matrix and removing redundant contact data allows easier comparison with other features. (*e*) Connecting arcs are a more effective visual encoding for spatial contacts, and a circular layout is generally more compact. Images *a-e* were made using Ensembl (123), Biodalliance (https://www.biodalliance.org , 124), JuiceBox (37), WashU EpiGenome Browser (38), and Rondo (http://rondo.ws, 40), respectively, and modified in Illustrator. Data in *c-e* is from Rao et al. (125).

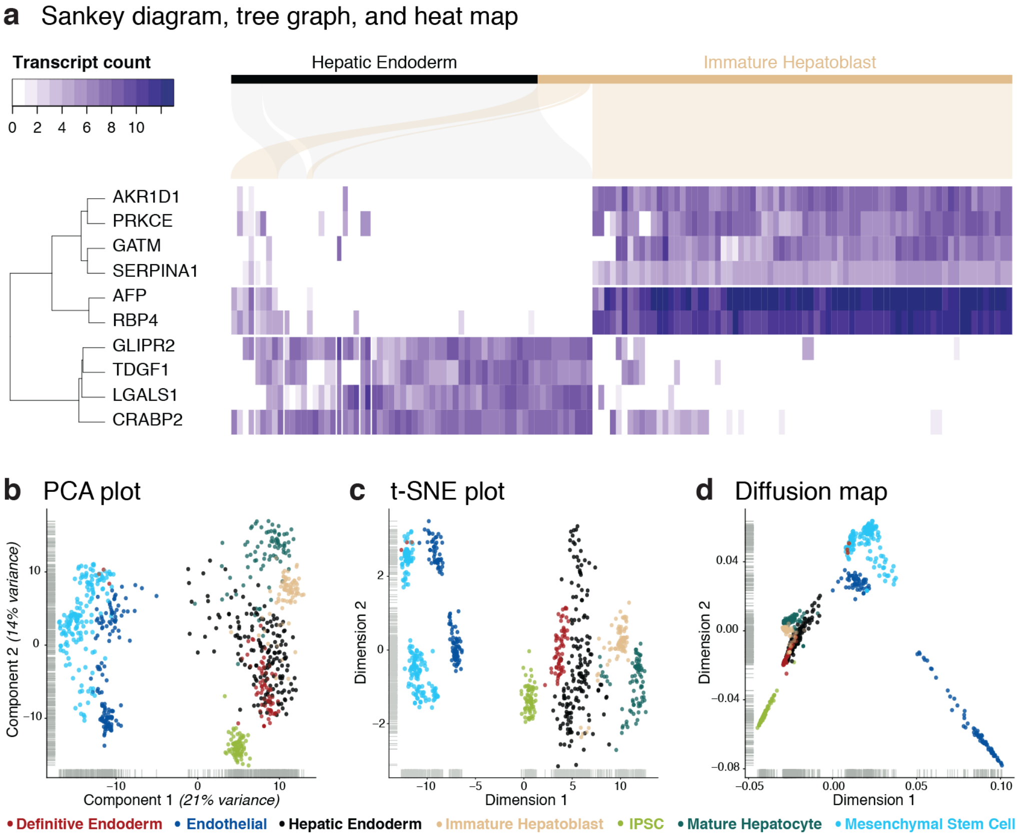


Figure 5

Visualizations of single-cell RNA-seq observations of liver bud development. (*a*) Clustered heatmap for top 10 differentially expressed genes in two cell types, indicated in the Sankey diagram with black and beige coloring. Absolute expression is encoded as saturation, and row and column position encodes genes and experimental conditions. Genes and cells with similar expression patterns are clustered to optimally order rows and columns. The cluster tree graph shows three distinct groups of expression behavior, and vertical space has been inserted to separate these sets of rows. The Sankey diagram highlights imperfect separation of the two cell types, and spaces have also been inserted to separate sets of differently behaving cells. Below, scatterplots show alternative views created by applying dimensionality reduction methods, each revealing different aspects of the full dataset. (*b*) Principal components analysis (PCA) groups most cell types, but does not resolve cells forming the definitive endoderm and the hepatic endoderm. (*c*) t-distributed stochastic neighborhood embedding (t-SNE, 126) provides more insight, revealing local similarities as well as overall variation in the dataset. However, t-SNE can be more difficult to apply, as it requires setting a manually adjustable parameter ('Perplexity', 127). (*d*) Diffusion maps (128, 129) model relationships between points in the dataset as a diffusion process that is then reduced to a lower-dimensional map. Here, successive developmental relationships between cells are revealed. Image *a* was made using R and D3.js (Sankey diagram). Images *b-d* were made using scater (130). All images were modified using Illustrator. Data in *a-d* is from Camp et al. (131), reanalyzed in R with read counts processed as described by Hemberg et al. (https://github.com/hemberg-lab/scRNA.seq.course).

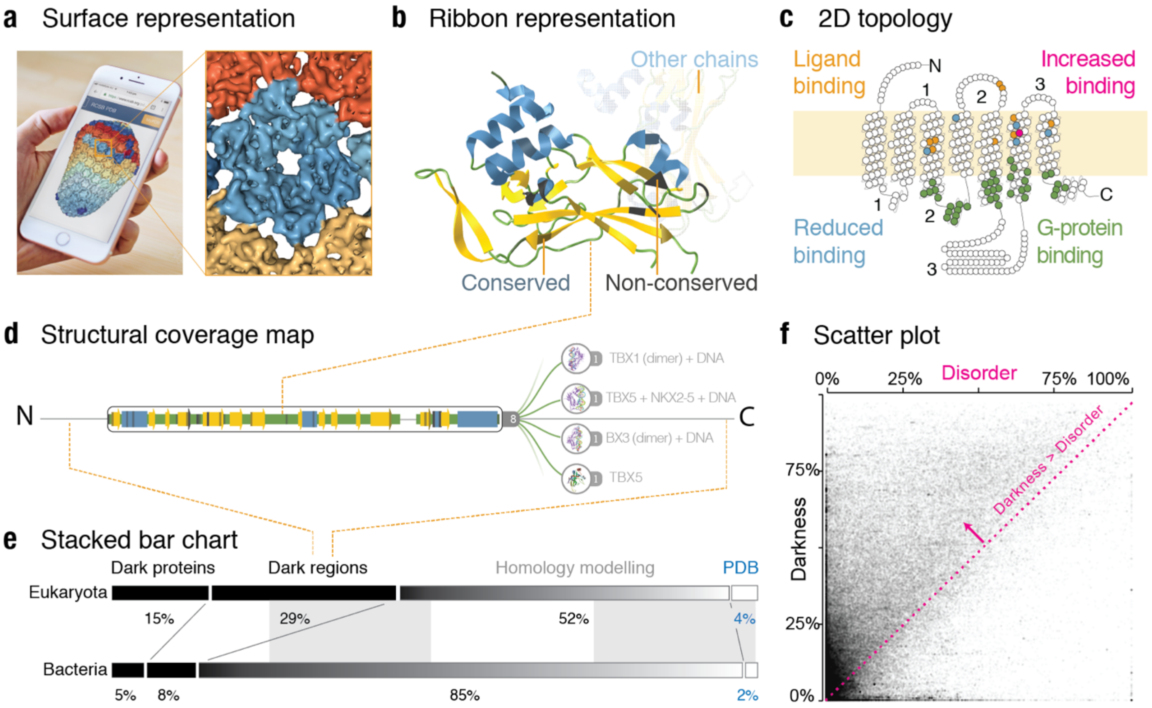


Figure 6

Visualizations of protein structure data. (*a*) Advances in web molecular graphics now allow structures with millions of atoms to be interactively explored on a smartphone. Surface (shown here) and space-fill representations are useful for overviewing the arrangement of individual proteins in larger assemblies. (*b*) For a detailed view of a single protein, ribbon representation (132) is useful, revealing how polypeptide chains fold in 3D; this is helped by linking to a sequence view (*d*). Using semi-transparency for other chains in the structure can provide supporting context without clutter. Highlighting conserved or non-conserved amino acid differences to the wildtype sequence of interest gives a visual indication of model reliability (133). Typically, many such differences occur in structures inferred via homology modeling; but they are also common in PDB structures, due to experimental limitations. (*c*) In special cases (e.g., GPCR transmembrane proteins, shown here), simplified 2D schematics can be used to show overall topology as well as details, such as loop regions or residues where mutations have large functional effects (encoded using numbering and coloring, respectively). (*d*) A schematic representation of a full-length, wildtype protein sequence, with coloring indicating regions with significant sequence similarity to structures in the PDB. Details on these matching structures can be revealed upon demand using a tree graph. On average, each protein sequence matches to ~200 PDB structures (62), but contains several 'dark' regions, with no detectable similarity to any known 3D structure (64). (*e*) Shows total fraction of protein residues that map to any PDB structure (either directly or via homology modelling); the remaining fraction ('dark proteome') is divided into dark regions (*d*) and dark proteins (where a single dark region spans an entire sequence) (64). To achieve moderate data density, stacked bar charts have been used, and axes replaced by two shaded regions (indicating 25%, 50%, and 75%). Connecting lines facilitate comparison. (*f*) A scatter plot of darkness versus disorder (134) for ~180,000 eukaryotic proteins. Point size, color, and transparency have been adjusted to reveal an unexpected, overall pattern (darkness > disorder for most proteins, indicating that much of the dark proteome is not explained by disorder). Due to high data density, subtle patterns are also revealed (e.g., horizontal streaks arising from related sequence families). Images in *a* were made using NGL Viewer (https://www.rcsb.org/pdb/ngl/ngl.do?pdbid=3J3Q, 135) with PDB 3J3Q (136), in *b* & *d* using Aquaria (http://aquaria.ws/O75333/4a04/, 62) and Photoshop with PDB 4A04 (137), in *c* with data from GPCRdb (138), in e & f using ggplot2 (116) with data from (139). All images were modified using Illustrator.

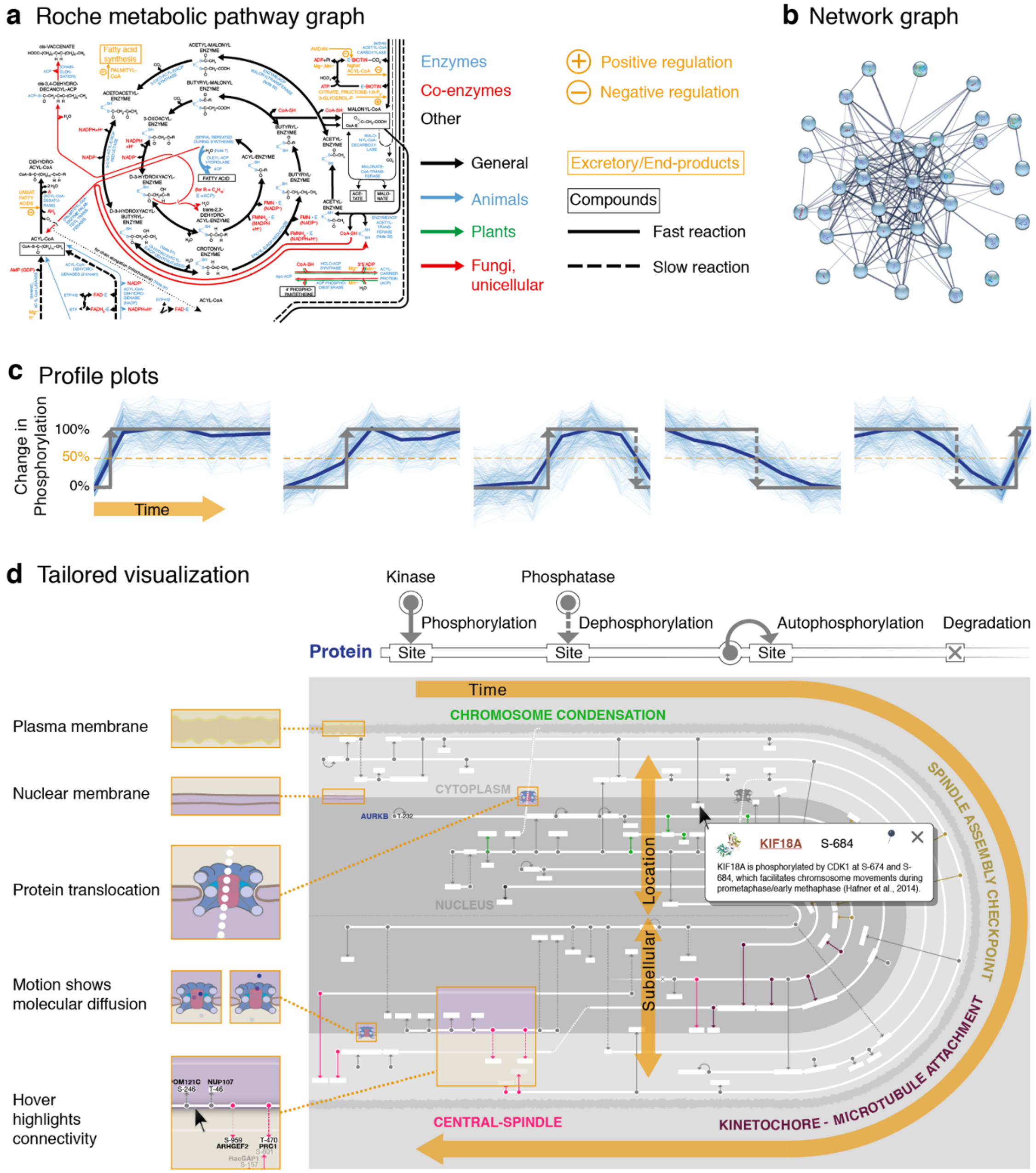


Figure 7

Pathway and network graphs of molecular systems. (*a*) Part of an exemplary tailored visualization; the full pathway shows causal flow involving ~3,000 reactions, plus supporting context (e.g., molecular structures). Note the effective use of visual channels (**Figure 2**): position and shape show reaction categories, while minimal coloring is used to show - without clutter - different versions of the pathway for four categories of organisms. (*b*) Network graphs created via spring-embedding are common (here, edge width encodes interaction confidence scores), but often too cluttered. This can be partly mitigated, e.g., via edge-bundling (76). (*c*) The first step in visualizing phosphoproteomics data is to identify clusters of phosphosites with similar time-profiles (thin light-blue lines). Clusters are modelled as a series of phosphorylation and dephosphylation events (solid and dashed arrows, respectively), each arising from a specific kinase or phosphatase (140) and occurring when the average phosphorylation (thick blue line) passes 50% (dotted line). (*d*) A tailored visualization for phosphoproteomics data where selected proteins are drawn as tracks in a 'circtangular' cellular landscape. Position encodes subcellular location (with translocations shown as excursions from the track layout) and the temporal ordering of phosphoevents, each drawn as an arrow connecting a kinase or phosphatase to its substrate site, indicated with residue numbering (not shown). Color hue shows events that perform coordinated functions. Texture is used show context, such as membranes. The online version has further interactive features (e.g., informative popups, motion, highlighting upon hover) that help researchers use these complex datasets to gain insight into cellular processes, such as insulin response (83) or mitosis (84). Image *a* was redrawn from http://biochemical-pathways.com/ (70), *b-d* were made using STRING (141), Matplotlib (115), and Minardo (https://minardo.org, 84), respectively, and modified using Illustrator.

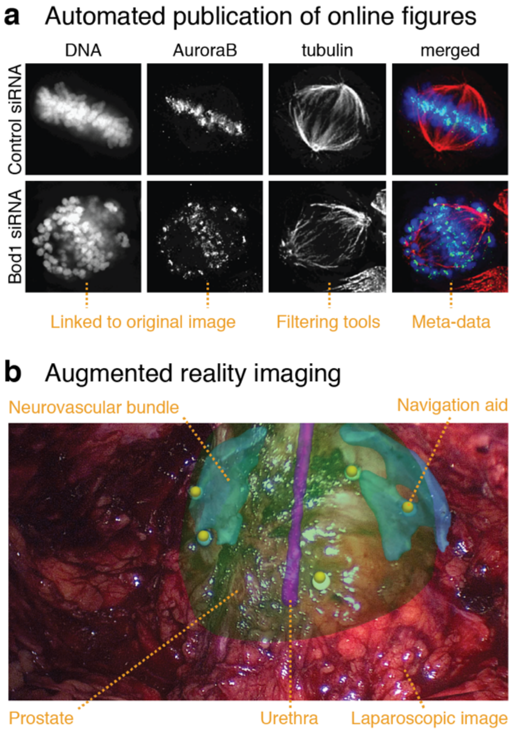


Figure 8

Cellular and tissue imaging. (*a*) Multi-part, annotated image created from, and linked to, raw datasets. By automating many routine manual tasks involved in creating well-formatted, publication-ready figures, tools such as OMERO.figure enable scientists to scale-up, easily creating figures with higher data density, and thus address more complex questions. Figure panels can be rendered dynamically from the original image data and automatically overlaid with timestamps and scale bars, avoiding potential human error. Such tools can document all steps from the original image files to the final figure, improving data integrity, organization, and provenance. (*b*) Augmented reality (AR) imaging in minimally-invasive surgery. Before the intervention, target and critical structures are segmented in a 3D planning image. At the beginning of surgery, artificial navigation aids (fiducials) are inserted into the target organ (here: prostate) and their 3D configuration is determined from a 3D intra-operative medical image (e.g. a 3D transrectal ultrasound image). The latter is fused with the pre-operative planning image using a 3D/3D registration algorithm. During surgery, the fiducials are continuously tracked in the 2D video images, and a 2D/3D registration algorithm (142) is used to find a transformation relating the endoscopic camera coordinate system with the image coordinate system of the 3D intra-operative modality. This enables the laparoscopic video image acquired during prostatectomy to be overlaid in real time with the prostate capsule and critical structures. Image *a* was made using OMERO.figure (http://figure.openmicroscopy.org/demo/#file/1, (89)) with data from Porter et al. (143); *b* using the Medical Imaging Tool Kit (144), with data from Simpfendörfer et al. (145).

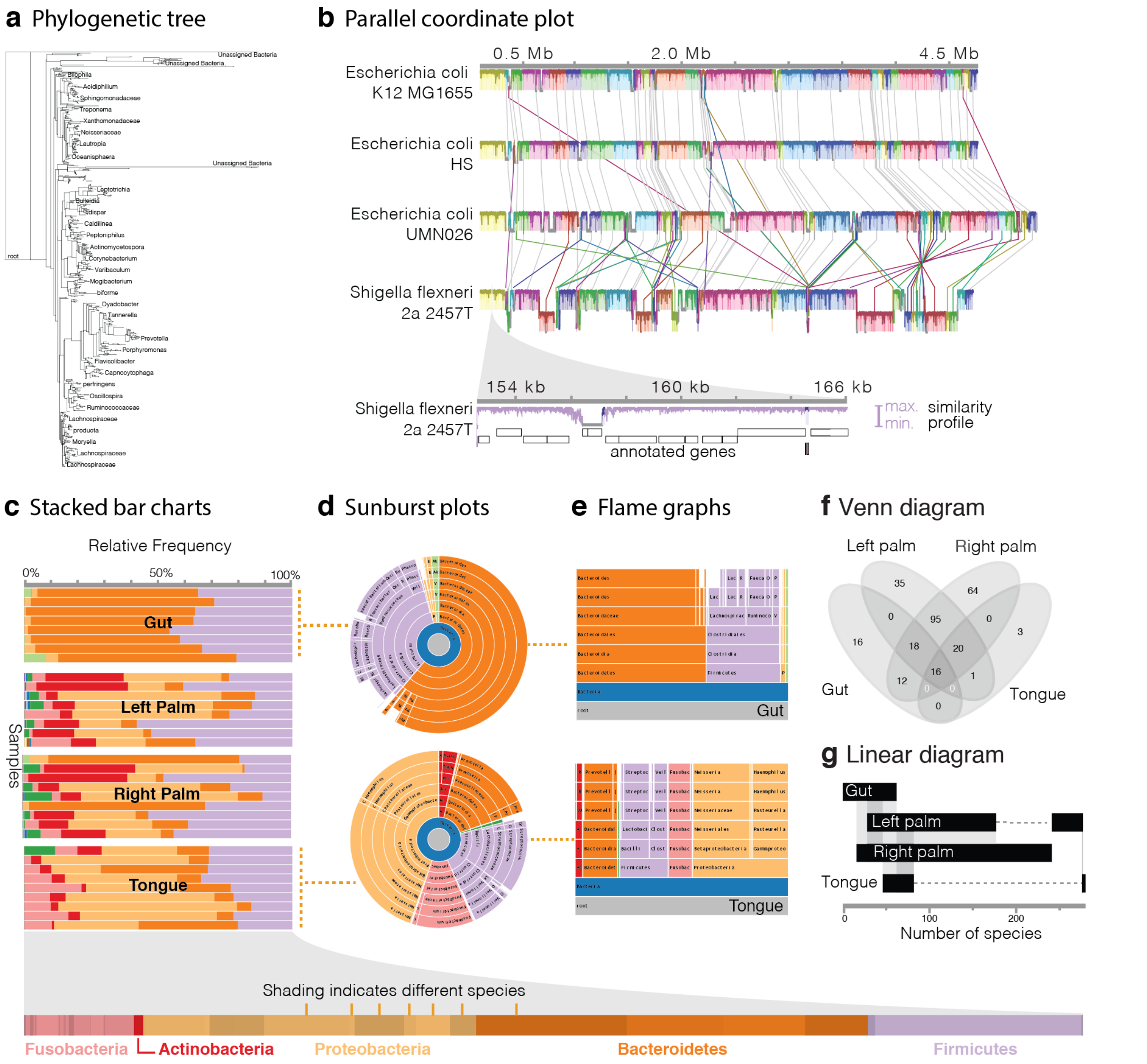


Figure 9

Phylogenetic, comparative genomics, and metagenomic data visualizations. (*a*) A phylogenetic tree showing evolutionary relationships and inferred operational taxonomic units (OTUs) for 16S amplicon sequencing data. Whilst commonly used, phylogenetic trees have drawbacks: closely related taxa become hard to resolve and as their number increase, topological relationships quickly become obscured, even with the application of semantic zooming (used here to omit overlapping labels on adjacent branches). The tree contains a total of 761 leaves, of which ~50 have a named OTU. (*b*) A multiple genome alignment viewed as a parallel coordinates plot, with a zoomed in region on a *Shigella flexneri* genome. Each genome is represented as a linear axis. Connecting lines between genomes indicate conserved regions. Lines that do not intersect with some genomes indicate horizontal gene transfer, and convergent and divergent lines correspond to gene duplication and inversion events between *Shigella* and the three *E. coli* strains. The zoomed region of the *Shigella* genome reveals regions of divergence in an otherwise conserved part of the alignment. Colors are assigned to aligned sections of each genome, and a similarity profile is overlaid as a line graph in a more saturated color. When genes are inverted in some organisms, the area is shown below the genome axis. (*c*) Species abundance (or beta-diversity) visualized as a stacked bar chart for a metagenomics analysis of microbiome samples taken from different parts of the body. Colors encode phyla of identified OTUs in samples. These charts are useful for comparing abundance across broad taxonomic levels (as shown here), but become too complex when used to show the ~280 species in each sample (zoomed region). (*d*) Sunburst plots showing beta-diversity for pooled samples from two body sites; whilst these accurately portray lineage relationships, it can be difficult to compare multiple plots. (*e*) Flame graphs encode taxonomic rank as height, and abundance as width, making it easier to compare plots and to see where taxonomic assignment is incomplete. (*f*) Species co-occurrence amongst samples from the four sites shown as a Venn diagram. Many tools offer advanced layout and shading models for Venn diagrams, but can result in plots that are not visually effective. (*g*)A linear diagram of the same data, using the *X* axis to show the number of co-occurring species between tissues, with gray vertical boxes highlighting intersections. Image *a* was made using Archaeopteryx (146), *b* using Mauve (147), *c* using QIIME2 (148), *d*-e using QuanTiTree (http://metasystems.riken.jp/visualization/quantitree/index.htm) *f* using the R ‘venn’ library (https://cran.r-project.org/web/packages/venn/index.html), and *g* using the Linear Diagram Generator (149). All images were modified using Illustrator. Data in *a,c-g* are from the ‘moving pictures’ dataset (150).