CANADA'S MICHAEL SMITH GENOME SCIENCES CENTRE

Getting Into Visualization of Large Biological Data Sets

20 IMPERATIVES OF INFORMATION DESIGN

Martin Krzywinski, Inanc Birol, Steven Jones, Marco Marra

BC Cancer Agency
CARE & RESEARCH
An agency of the Provincial Health Services Authority

ENSURE LEGIBILITY AND FOCUS ON THE MESSAGE

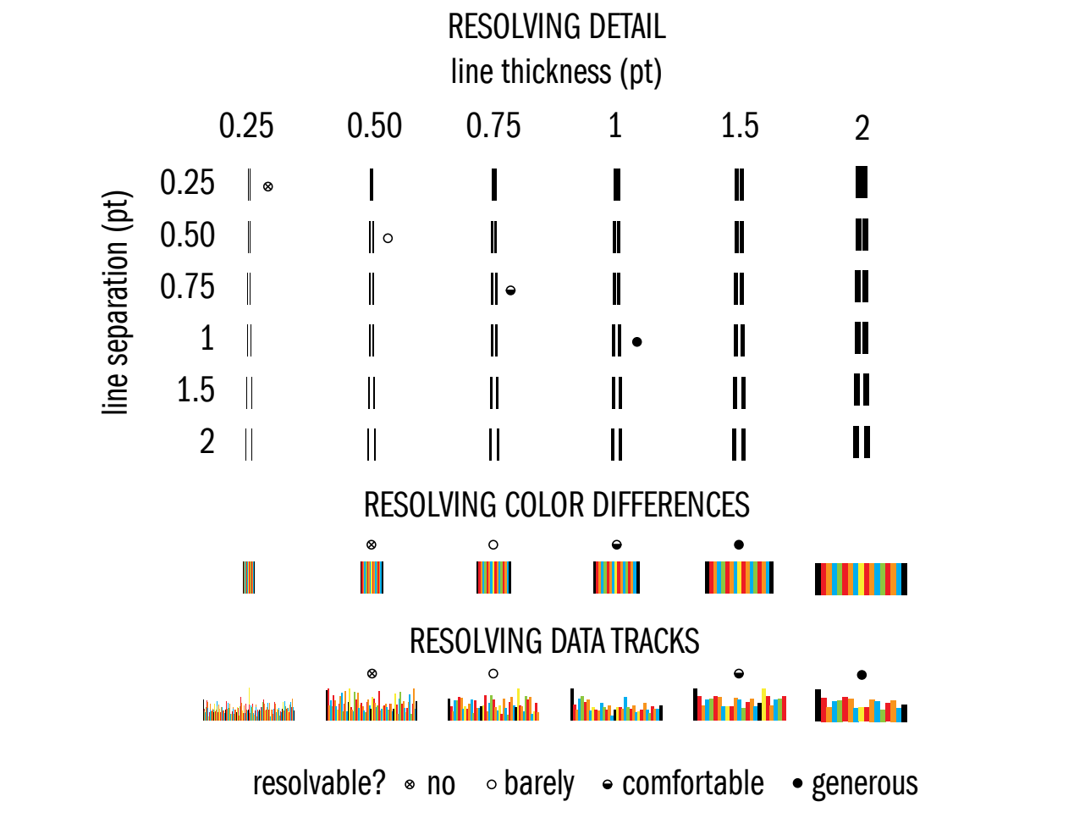
Create legible visualizations with a strong message. Make elements large enough to be resolved comfortably. Bin dense data to avoid sacrificing clarity.

Distinguish between exploration and communication.

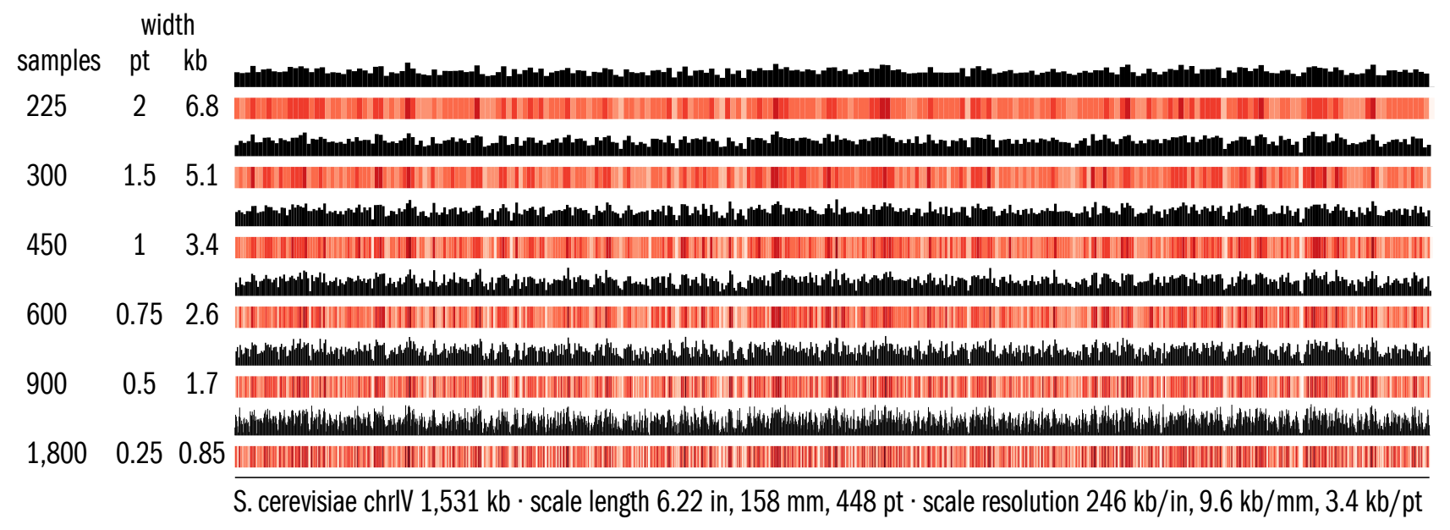
Use exploratory tools (e.g. genome browsers) to discover patterns and validate hypotheses. Avoid using screenshots from these applications for communication – they are typically too complex and cluttered with navigational elements to be an effective static figure.

Do not exceed resolution of visual acuity.

Our acuity is ~50 cycles/degree or about $^1/_{200}$ (0.3 pt) at 10 inches. Ensure the reader can comfortably see detail by limiting resolution to no more than 50% of acuity. Where possible, elements that require visual separation should be at least 1 pt part.



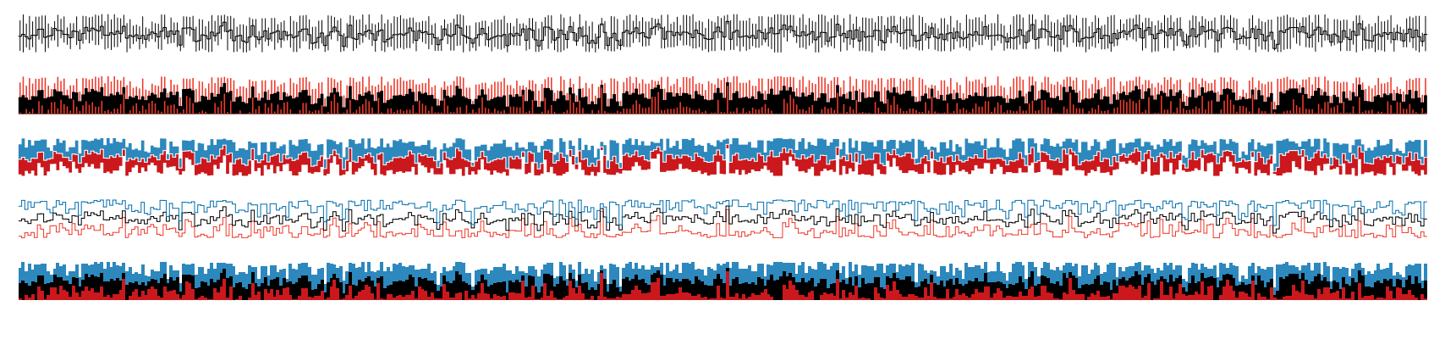
Visual acuity limits the perception of color and spatial differences. For standard reading distance, we can comfortably resolve objects separated by 1 pt. For larger viewing distances (e.g. this poster), this limit is proportionately higher. The standard unit of distance in print is 1 pt = 1/72 inch.



Rendering human chromosome 1 (249 Mb) with 450 divisions provides a resolution of 550 kb, which is larger than 98% of all human genes. To depict 50% of genes without magnification we would need to use 15 kb bins, which would have to be 1/16,600 of the figure (0.0004 in, 10 microns)!

Use no more than ~500 scale intervals.

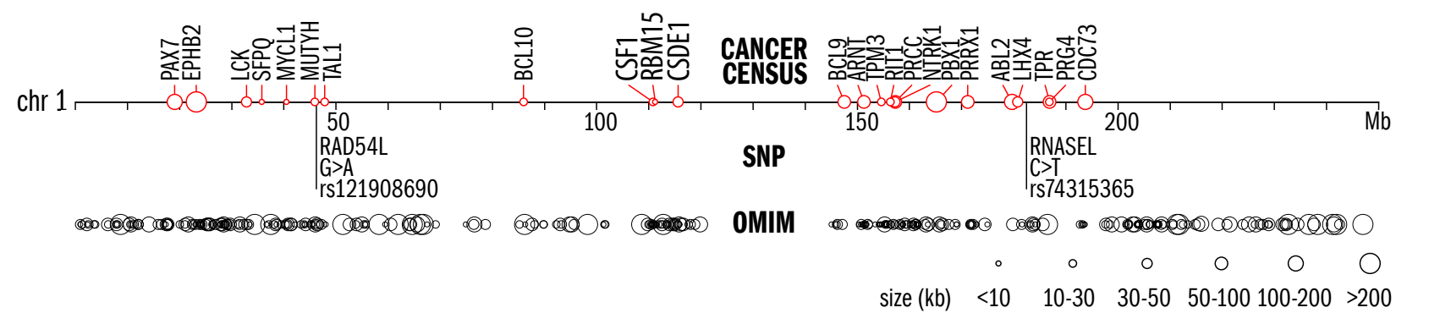
Ensure data elements are at least 1 pt on a two-column Nature figure (6.22 in), 4 pixels on a 1920 horizontal resolution display, or 2 pixels on a typical LCD projector. These restrictions become challenges for large genomes.



Approaches to encoding min/avg/max values of downsampled data. In the top hi-low trace, the vertical bars are perceived as a separate layer and effectively show variance without obscuring trends in the average.

Show variation with statistics.

Data on large genomes must be downsampled. Depict variation with min/max plots and consider hiding it when it is within noise levels. Help the reader notice significant outliers.



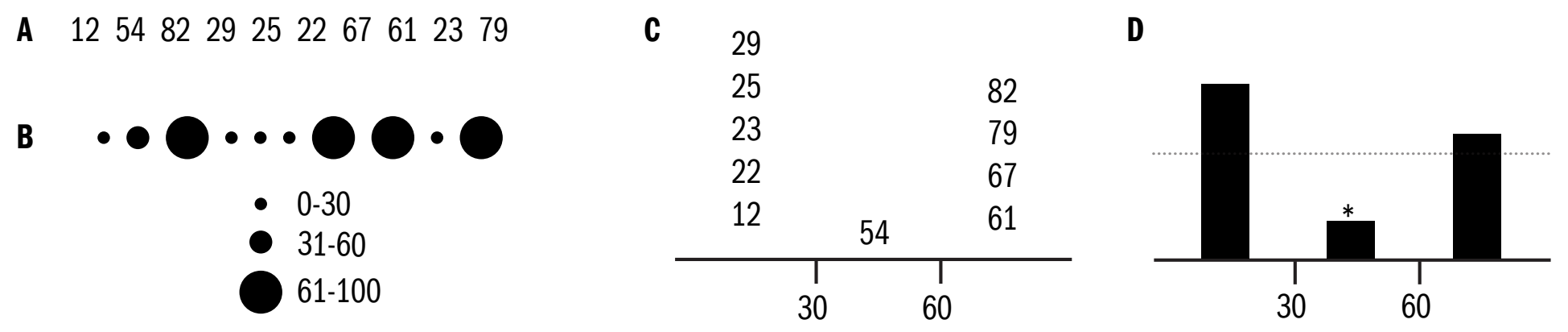
When drawing the position and size of densely packed genes, encode the gene's size using a non-linear mapping. When the number of data values is large, such as in the OMIM gene track, hollow glyphs are effective. For even greater number of points, a density map is preferred.

Do not draw small elements to scale.

Map size of elements onto clearly legible symbols. Legibility and clarity are more important than precise positioning and sizing. Discretize sizes and positions to facilitate making meaningful comparisons.

Aggregate data for focused theme.

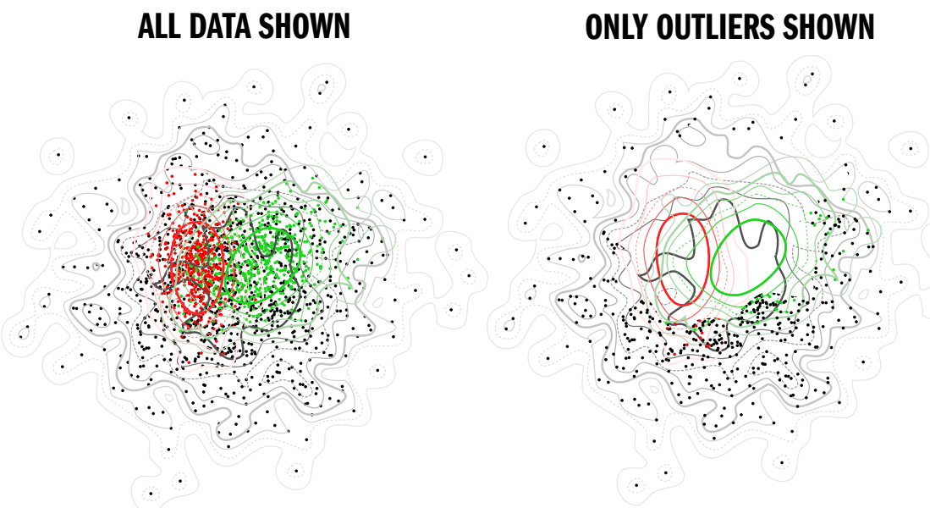
A strong visual message has no uncertainty in its interpretation. Focus on a single theme by aggregating unnecessary detail.



What is communicated? (A) The raw data imparts no clear message.(B). Binning indicates ranges, not individual values, are important. (C). Frequency distribution suggests that there is a shortage of medium-sized values. (D) Individual data points can be removed to emphasize trend and significance.

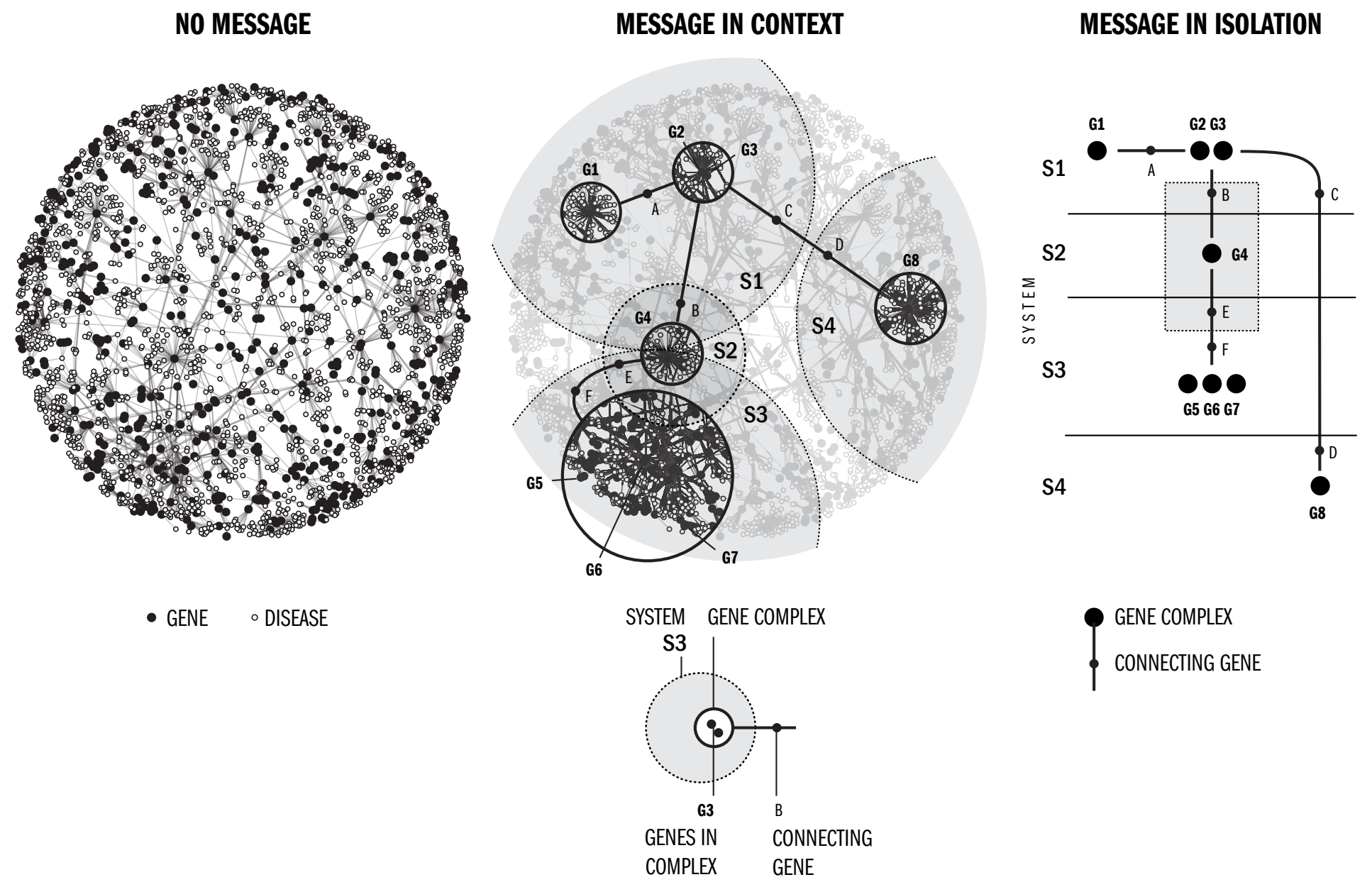
Show density maps and outliers.

Establishing context is helpful when emergent patterns in the data provide a useful perspective on the message. When data sets are large, it is difficult to maintain detail in the context layer because the density of points can visually overwhelm the area of interest. In this case, consider showing only the outliers in the data set.



Consider whether showing the full data set is useful.

The reader's attention can be focused by increasing the salience of interesting patterns. Other complex data sets, such as networks, are shown more effectively when context is carefully edited or even removed.



USE EFFECTIVE VISUAL ENCODINGS TO ORGANIZE INFORMATION.

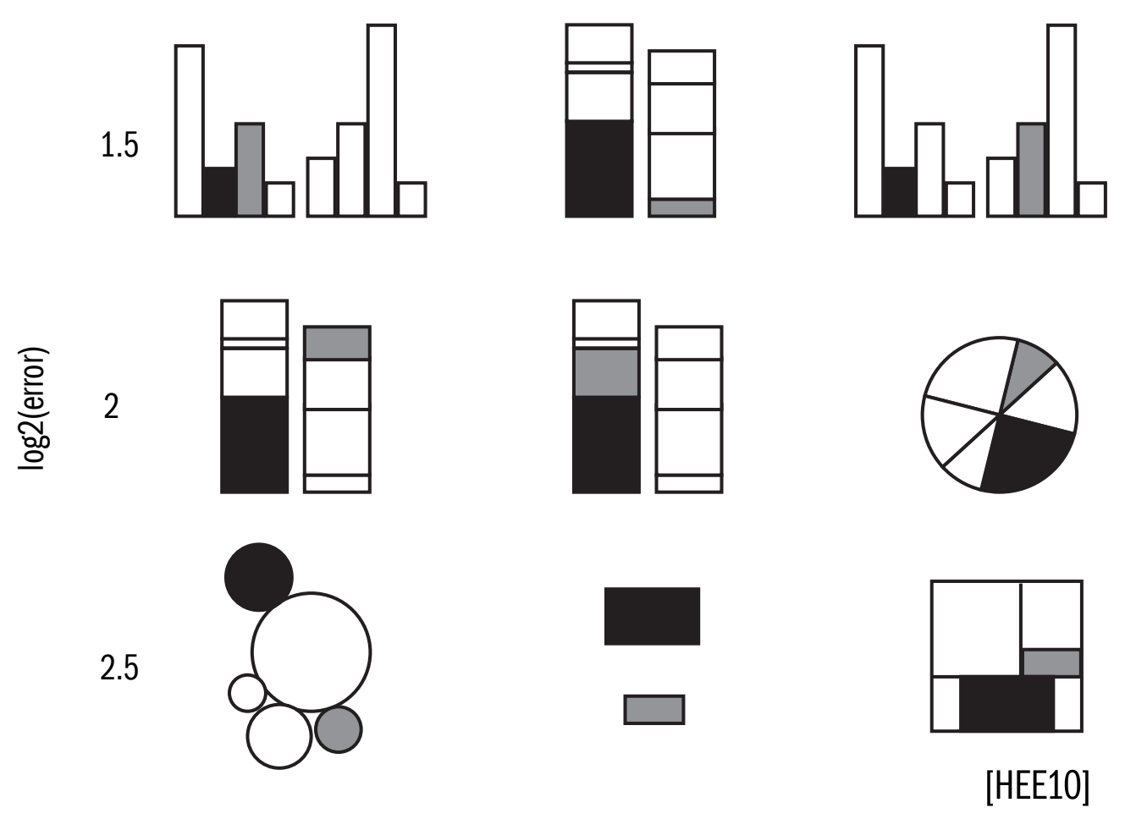
Match the visual encoding to the hypothesis. Use encodings specific and sensitive to important patterns. Dense annotations should be independent of the core data in distinct visual layers.

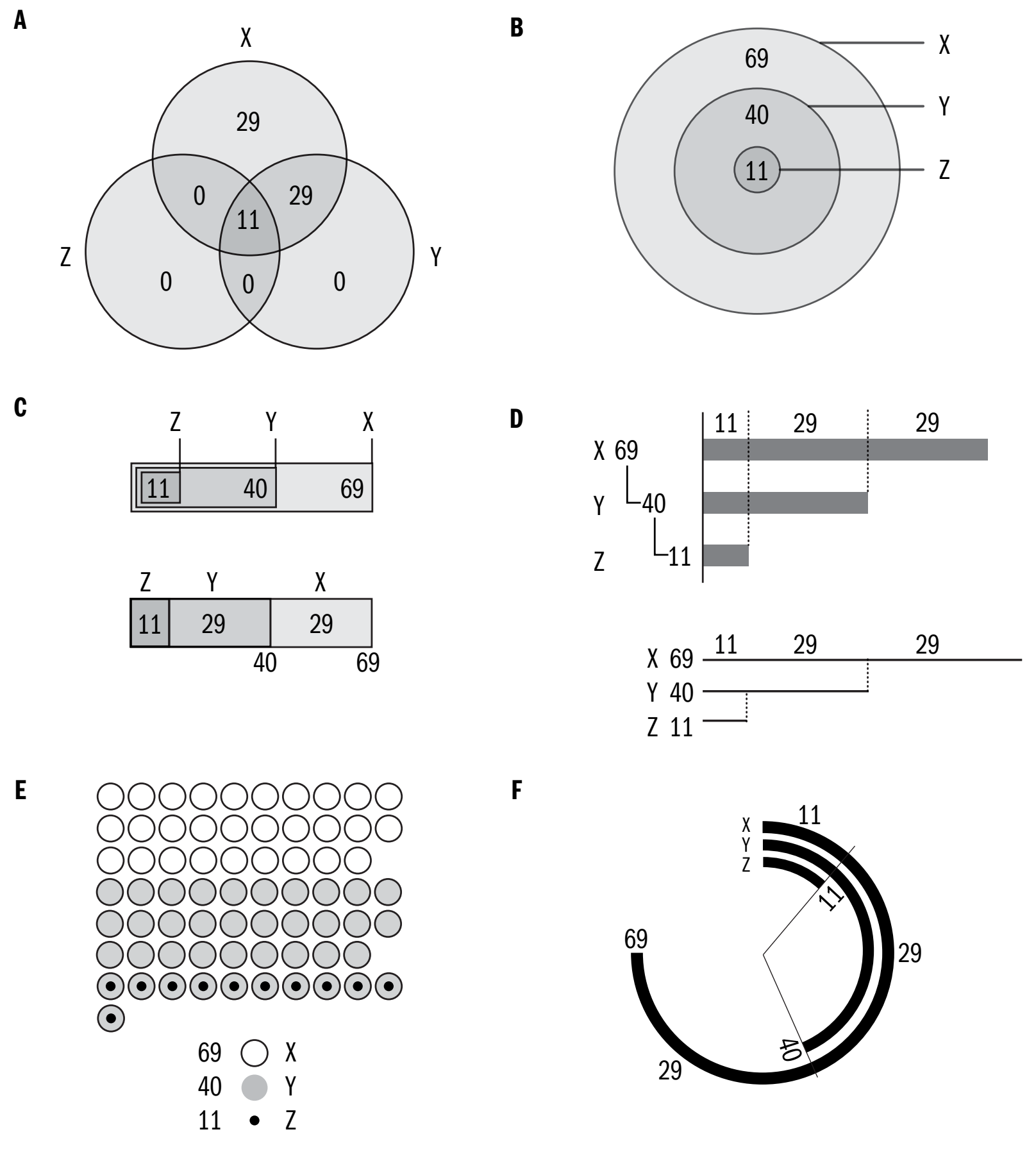
Use the simplest encoding.

Choose concise encodings over elaborate ones.

Help the reader judge accurately.

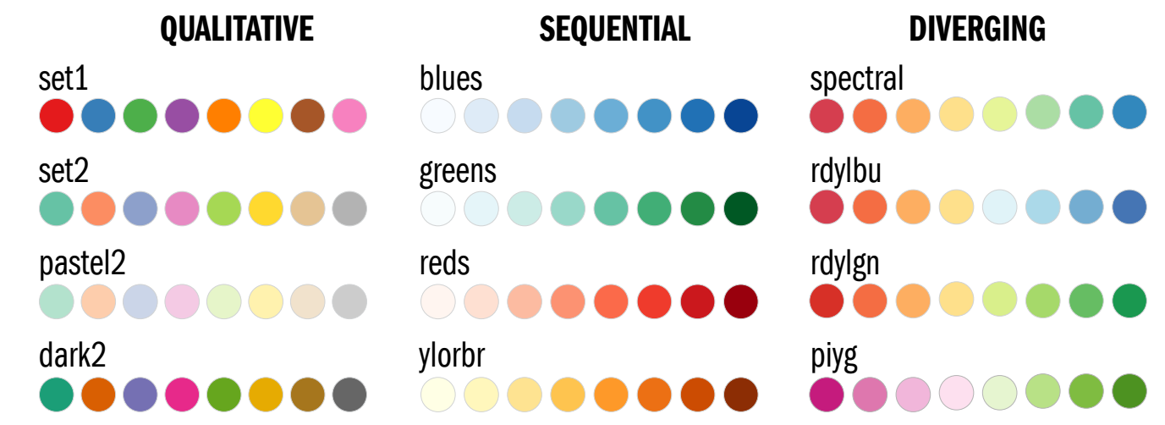
Accuracy and speed in detecting differences in visual forms depends on how information is presented. We judge relative lengths more accurately than areas, particularly when elements are aligned and adjacent. Our judgment of area is poor because we use length as a proxy, which causes us to systematically underestimate.





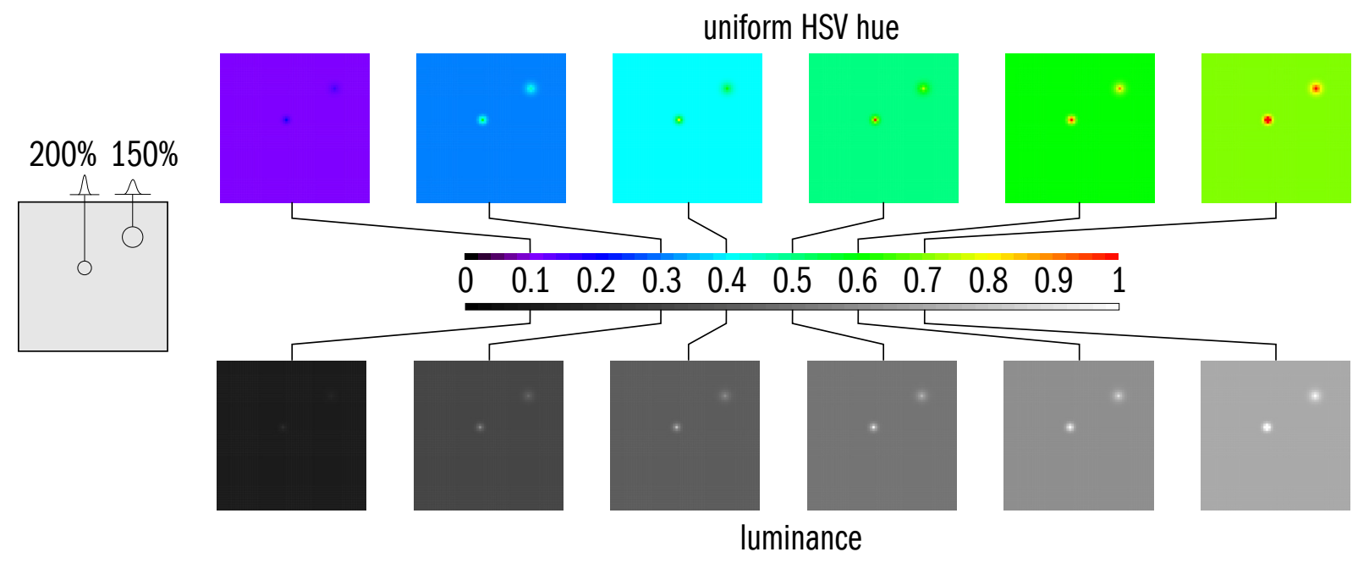
Use perceptual palettes.

Selecting perceptually favorable colors is difficult because most software does not support the required color spaces. Brewer palettes [BRE11] exist for the full range of colors to help us make useful choices. Qualitative palettes have no perceived order of importance. Sequential palettes are suitable for heat maps because they have a natural order and the perceived difference between adjacent colors is constant. Twin hue diverging palettes, are useful for two-sided quantitative encodings, such as immunofluorescence and copy number.



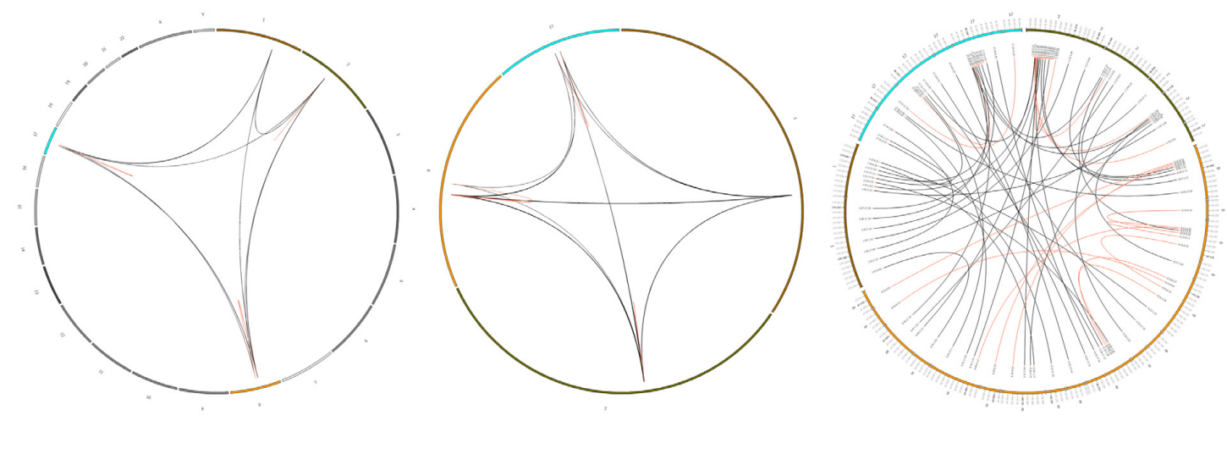
Never use hue to encode magnitude.

Hue does not communicate relative change in values because we perceive hue categorically (blue, green, yellow, etc). Changes within one category have less perceptual impact than transitions between categories. For example, variations across the green/yellow boundary are perceived to be larger than variations across the same sized hue interval in other parts of the spectrum.



Crop scale to reveal fine structure in data.

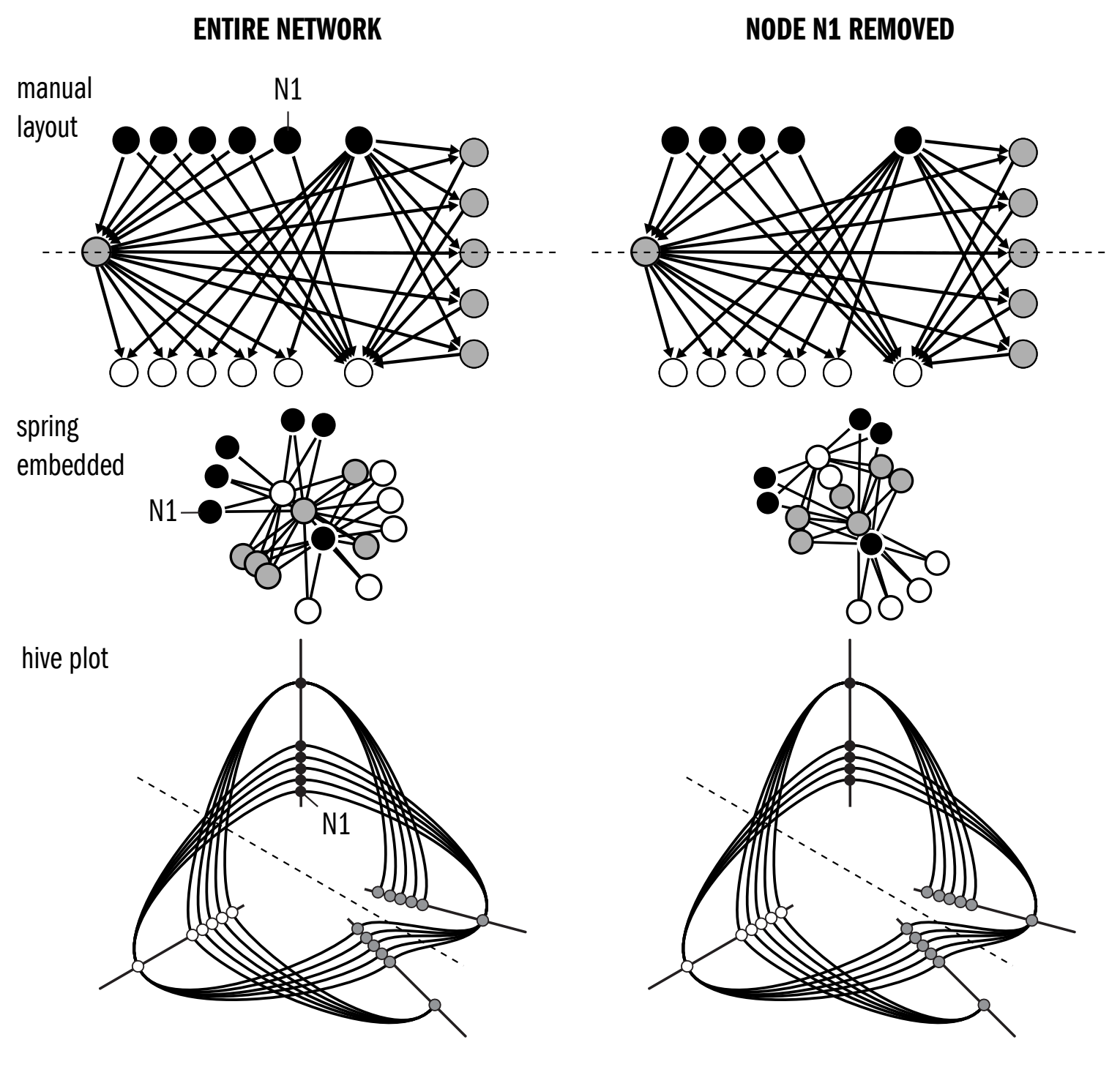
Biological data sets are typically high-resolution (changes at base pair level can meaningful), sparse (distances between changes are orders of magnitude greater than the affected areas) and connect distant regions by adjacency relationships (gene fusions and other rearrangements). It is difficult to take these properties into account on a fixed linear scale, the kind used by traditional genome browsers. To mitigate this, crop and order axis segments arbitrarily and apply a scale adjustment to a segment or portion thereof.



Circos image of patterns in densely distributed breakpoint positions between chromosomes 1, 2, 9 and 17 are emphasized when scale is zoomed. (A) full genome (B) chrs 1, 2, 9, 17 (C) 31 regions on chrs 1, 2, 9, 17 totaling 223 kb [KRZ09].

Use encodings that are robust and comparable.

In addition to being transparent and predictable, visualizations must be robust with respect to the data. Changes in the data set should be reflected by proportionate changes in the visualization. Be wary of force-directed network layouts, which have low spatial autocorrelation. In general, these are neither sensitive nor specific to patterns of interest.



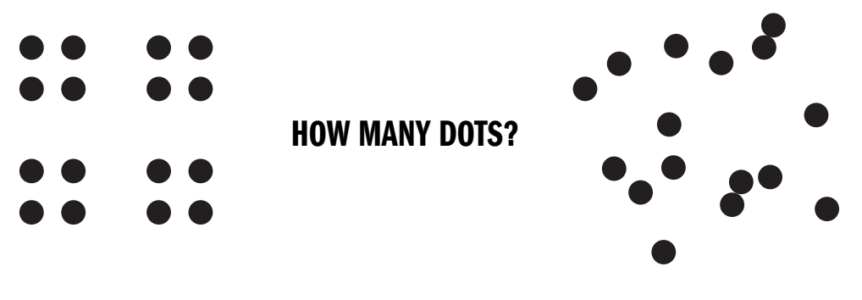
Hive Plots are a robust and quantitative visual form of networks based on meaningful properties chosen by the user. Unlike in a conventional layout, the removal of a node from a network can be easily spotted in a hive plot [KRZ11].

USE EFFECTIVE DESIGN PRINCIPLES TO EMPHASIZE AND COMMUNICATE PATTERNS.

Well-designed figures illustrate complex concepts and patterns that may be difficult to express concisely in words. Figures that are clear, concise and attractive are effective – they form a strong connection with the reader and communicate with immediacy. These qualities can be achieved with methods of graphic design, which are based on theories of how we perceive, interpret and organize visual information.

Reduce unnecessary variation.

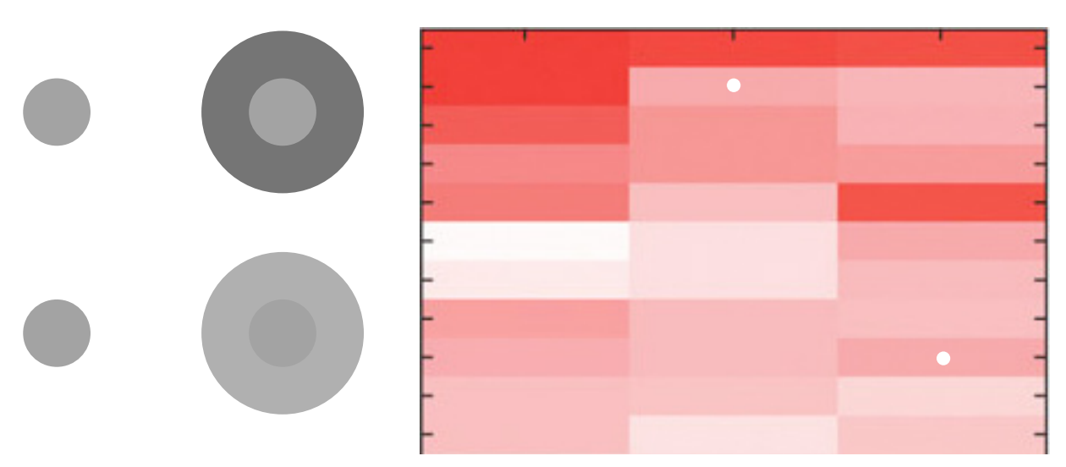
The reader does not know what is important in a figure and will assume that any spatial or color variation is meaningful. The figure's variation should come solely from data or act to organize information.



You can count the objects on the left almost instantly, without conscious effort. This is called pre-attentive cognition. This example uses the concept of proximity to partition the dots into pre-attentive groupings. Elements can be effectively organized by reducing spatial variation, here achieved by symmetrical layout and two levels of spacing,

Be aware of the luminance effect.

Color is a useful encoding – the eye can distinguish about 450 levels of gray, 150 hues, and 10-60 levels of saturation, depending on the color – but our ability to perceive differences varies with context. Adjacent tones with different luminance values can interfere with discrimination, in interaction known as the luminance effect.

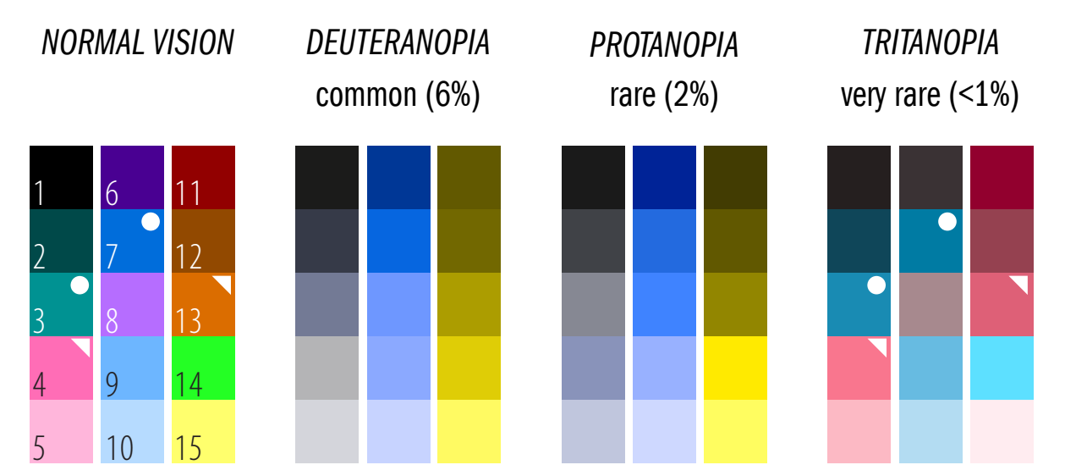


[GRE10] [WON10]

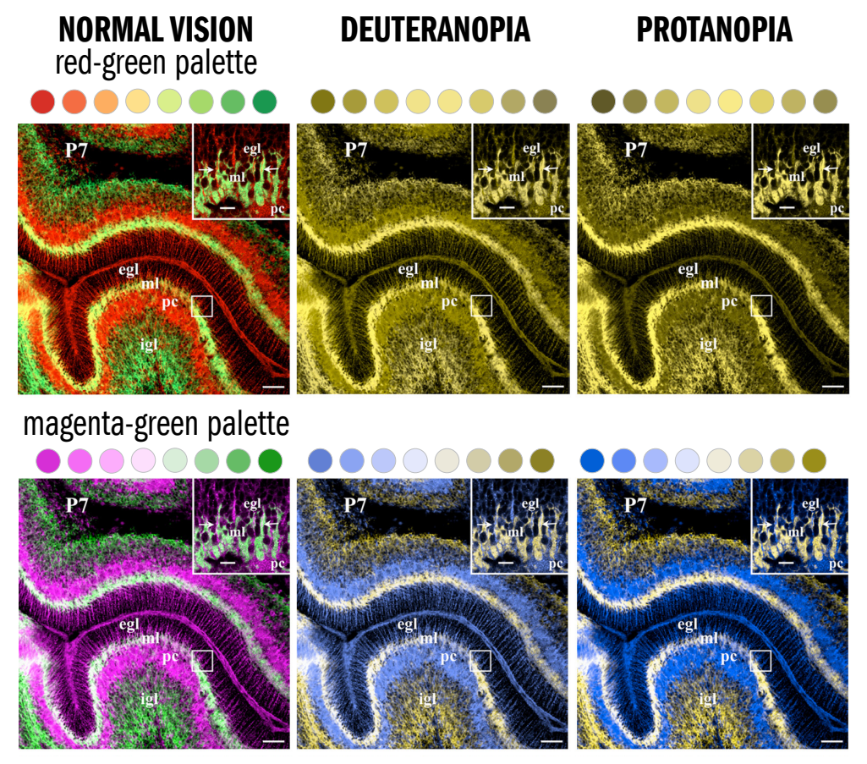
Be aware of color blindness.

In an audience of 8 men and 8 women, chances are 50% that at least one has some degree of color blindness. Use a palette that is color-blind safe. In the palette below the 15 colors appear as 5-color tone progressions to those with color blindness. Additional encodings can be achieved with symbols or line thickness.

15-COLOR PALETTE FOR COLOR BLINDNESS



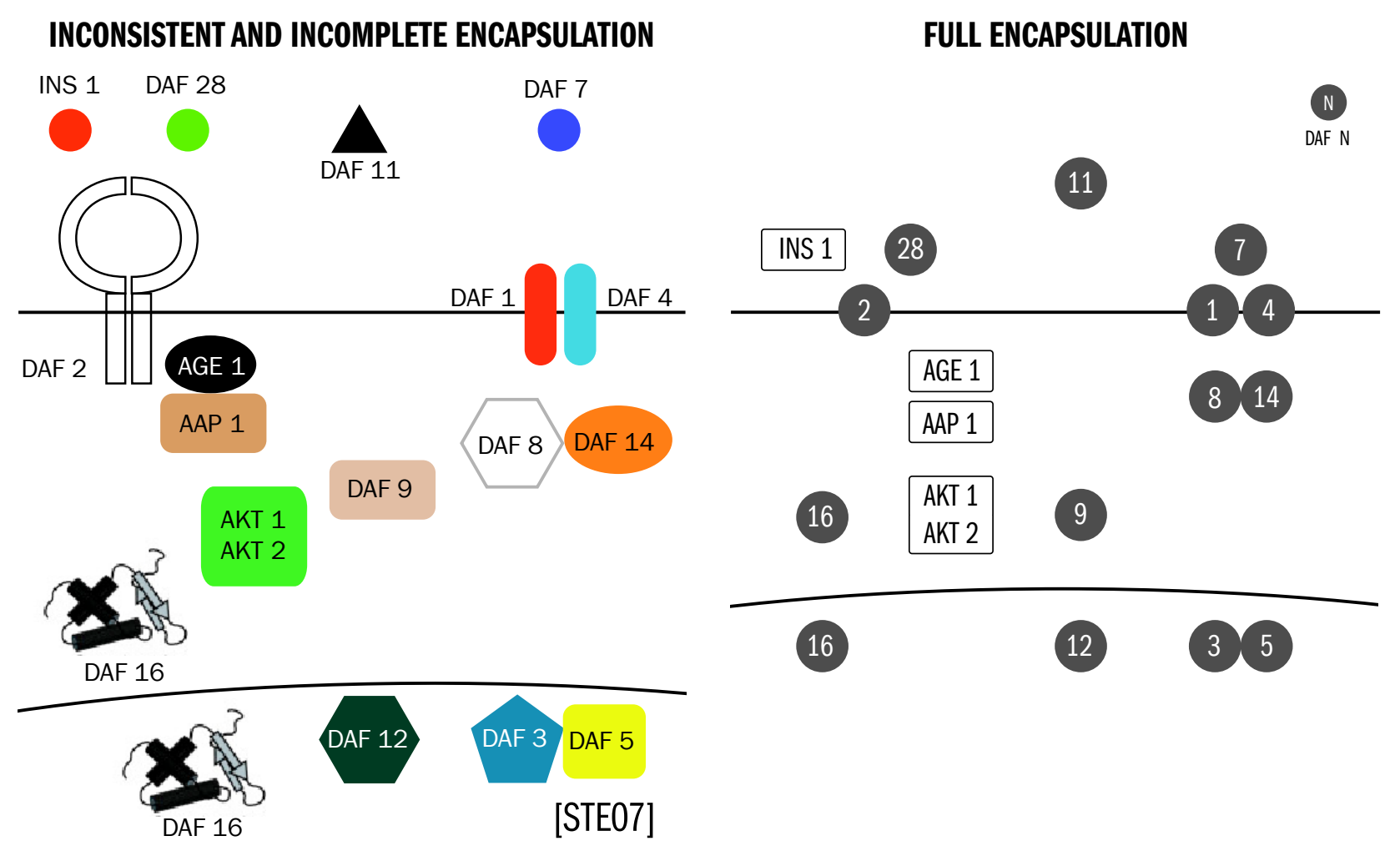
RGB											
1	0	0	0	6	73	0	146	11	146	0	0
2	0	73	73	7	0	109	219	12	146	73	0
3	0	146	146	8	182	109	255	13	219	209	0
4	255	109	182	9	109	182	255	14	36	255	36
5	255	182	119	10	182	219	255	15	255	255	109



Some Brewer palettes, such as the pink-yellow-green diverging palette, are color-blind safe. Use them for images where red/green discrimination is critical for comprehension.

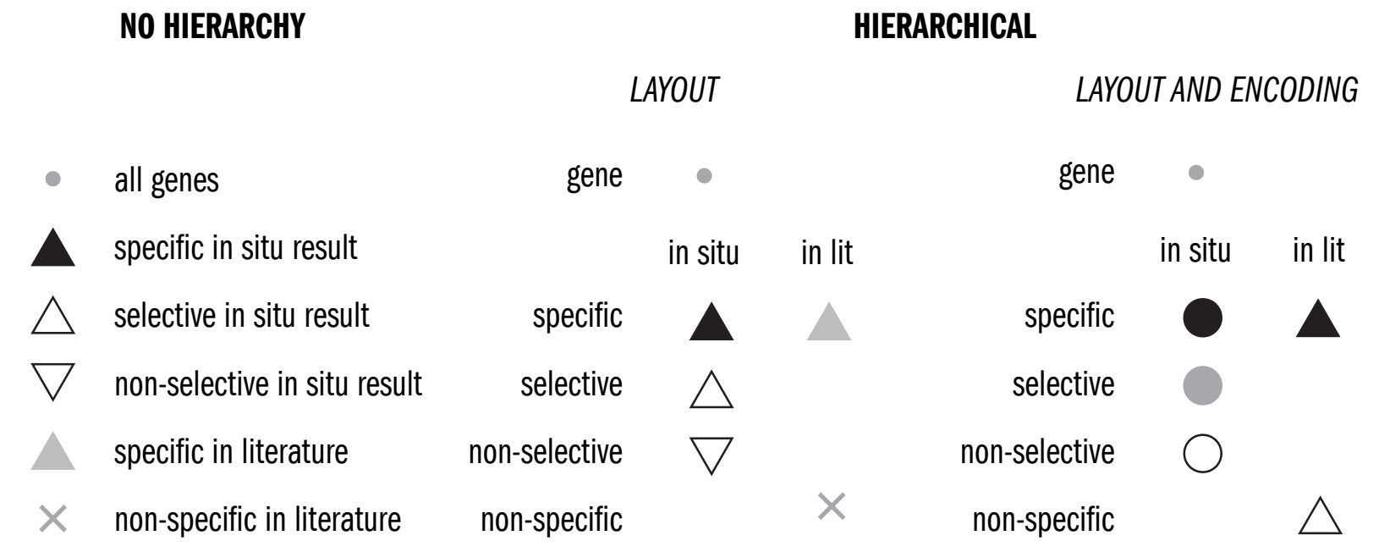
Encapsulate details.

Including details not relevant to the core message of the figure can create confusion. Encapsulation should be done to the same level of detail and to the simplest visual form. Duplication in labels should be avoided unless required to preserve semantic forms.



Respect natural hierarchies.

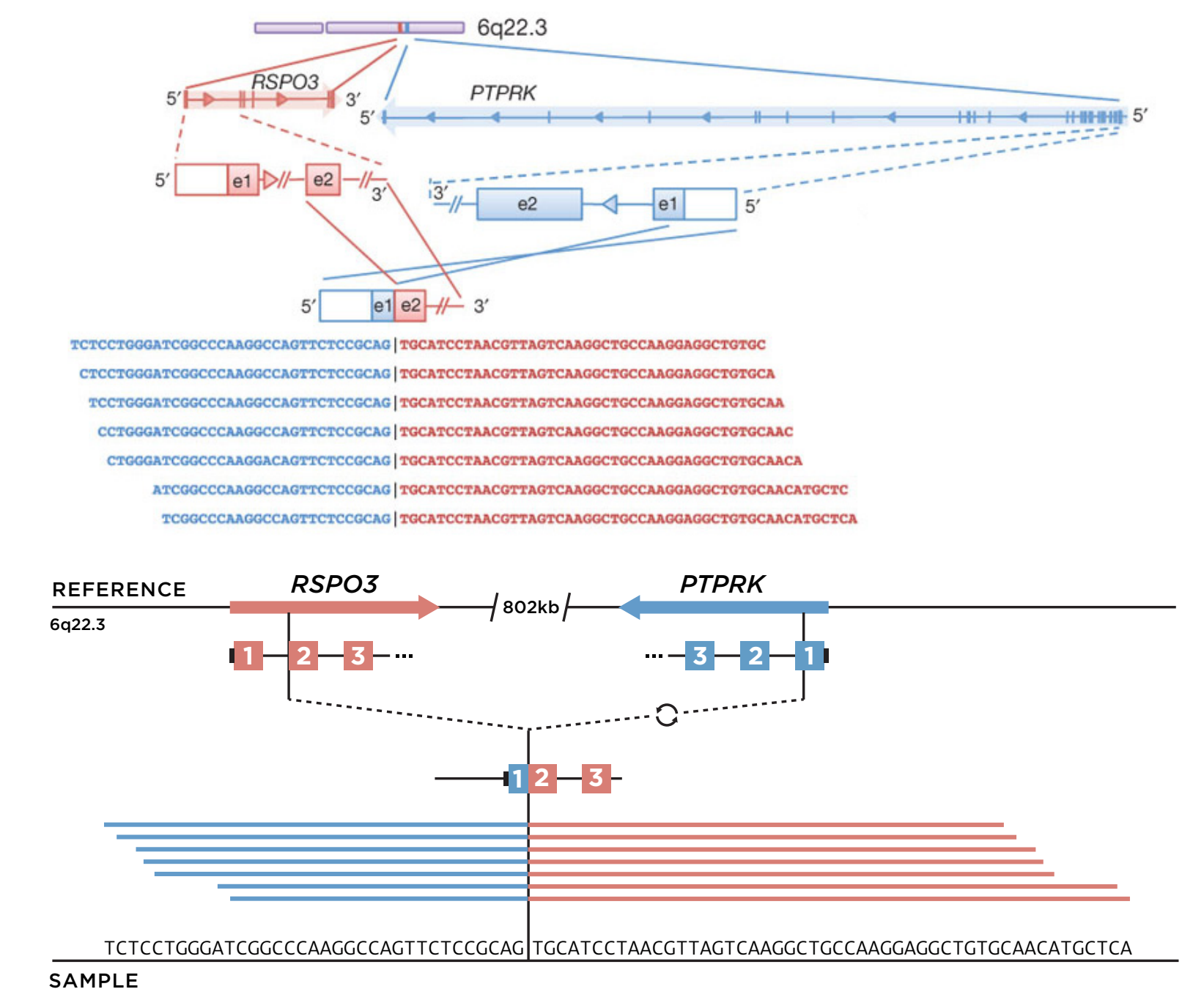
When the data set embodies a natural hierarchy, use an encoding that emphasizes it clearly and memorably. The use hierarchy in layout (e.g. tabular form) and encoding can significantly improve a muddled figure.



Adapted from [NEL03].

Use consistent alignment. Center on theme.

Establish equivalence using consistent alignment. Awkward callouts can be avoided if elements are logically placed.


6q22.3
RSPO3
PTPRK
5'
3'
e1
e2
REFERENCE
6q22.3
802kb
1
2
3
SAMPLE
TCTCCTGGGATCGGCCCAAGGCCAGTTCTCCGCAG TGCATCCTAACGTTAGTCAAGGCTGCCAAGGAGGCTGTGCAACATGCTCA

Complex information can be organized by consistent alignment. Notice how placing the gene fusion product in the center emphasizes the subject of the figure. Explaining the concept of gene fusion does not require the internal structure of the two genes, whose size and exon count/distribution should be removed to improve clarity. Original figure from [SES12].

[BRE11] C Brewer (2011) Color Brewer. http://www.colorbrewer.org
[GRE10] HE Grecco et al. (2010) In situ analysis of tyrosine phosphorylation networks by FLIM on cell arrays. Nat Methods 7: 467-472.
[HEE10] J Heer et al. (2010) Crowdsourcing graphical perception: using mechanical turk to assess visualization design. Proceedings of the 28th international conference on Human factors in computing systems. Atlanta, Georgia, USA: ACM. pp. 203-212.
[KRZ09] M Krzywinski et al. (2009) Circos: an information aesthetic for comparative genomics. Genome Res, vol. 19, pp. 1639-45
[KRZ11] M Krzywinski et al. (2011) Hive plots–rational approach to visualizing networks, Brief Bioinform, Dec 9 2011.
[NEL03] S Nelander et al. (2003) Prediction of cell type-specific gene modules: identification and initial characterization of a core set of smooth muscle-specific genes. Genome Res 13: 1838-1854.
[SES12] S Seshagiri et al. (2012) Recurrent R-spondin fusions in colon cancer. Nature 488: 660-664.
[STE07] SE Von Stetina et al. (2007) Cell-specific microarray profiling experiments reveal a comprehensive picture of gene expression in the C. elegans nervous system. Genome Biol 8: R135.
[WON10] B Wong (2010) Points of view: Color coding. Nat Methods 7: 573.

Content of this poster is drawn from my book chapter *Visualization Principles for Scientific Communication* (Martin Krzywinski & Jonathan Corum) in the upcoming open access Cambridge Press book **Visualizing biological data - a practical guide** (Seán I. O'Donoghue, James B. Procter, Kate Patterson, eds.), a survey of best practices and unsolved problems in biological visualization. This book project was conceptualized and initiated at the Vizbi 2011 conference.

If you are interested in guidelines for data encoding and visualization in biology, see the *Points of View* column by Bang Wong in Nature Methods.

v2 16 oct 2012
download poster