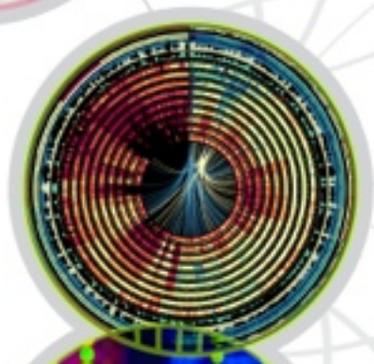
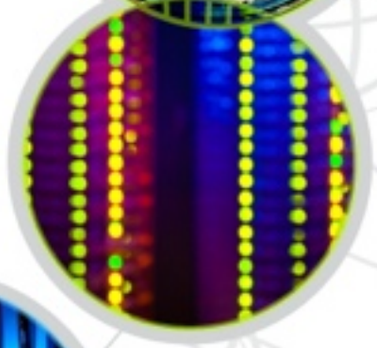
GENOMICS

INNOVATION

CANADA'S MICHAEL SMITH
GENOME
SCIENCES
CENTRE





INFORMATICS

SEQUENCING

COMPUTING

SCHLOSS DAGSTUHL
Leibniz-Zentrum für Informatik

About Dagstuhl | Program | Publications

You are here: **Program** » **Calendar** » Seminar Homepage

- Dagstuhl Seminars
- Dagstuhl Perspectives
- GI-Dagstuhl Seminars
- Events
- Research Guests
- Calendar
  - Seminars
  - Events

http://www.dagstuhl.de/12372

09.09.12 — 14.09.12, Seminar 12372

Biological Data Visualization

Organizers

Carsten Goerg (University of Colorado, US)
Lawrence Hunter (University of Colorado, US)
Jessie Kennedy (Edinburgh Napier University, GB)
Sean O'Donoghue (CSIRO - North Ryde, AU)
Jarke J. Van Wijk (TU Eindhoven, NL)

http://www.dagstuhl.de/en/program/calendar/semhp/?semnr=12372

This is an annotated version of the talk I gave at Schloss Dagstuhl at the Biological Data Visualization seminar.

visualization

communicating, clearly

annotated version

martin krzywinski

bc cancer research center
vancouver canada

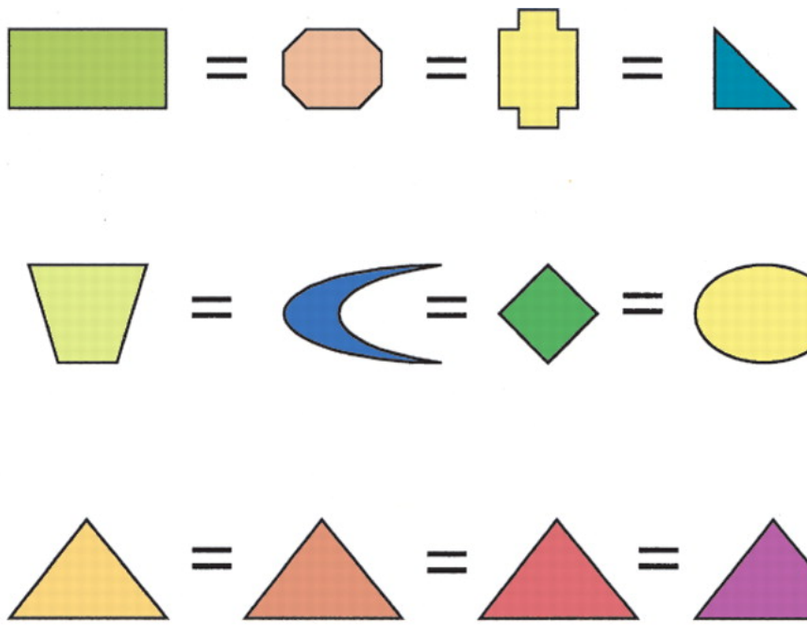
CREATE NECESSARY AND HELPFUL VISUALS

lushlush.livejournal.com/190093.html

GREATER ACCURACY YIELDS NO IMPROVEMENT

lushlush.livejournal.com/190093.html?thread=967053#t967053

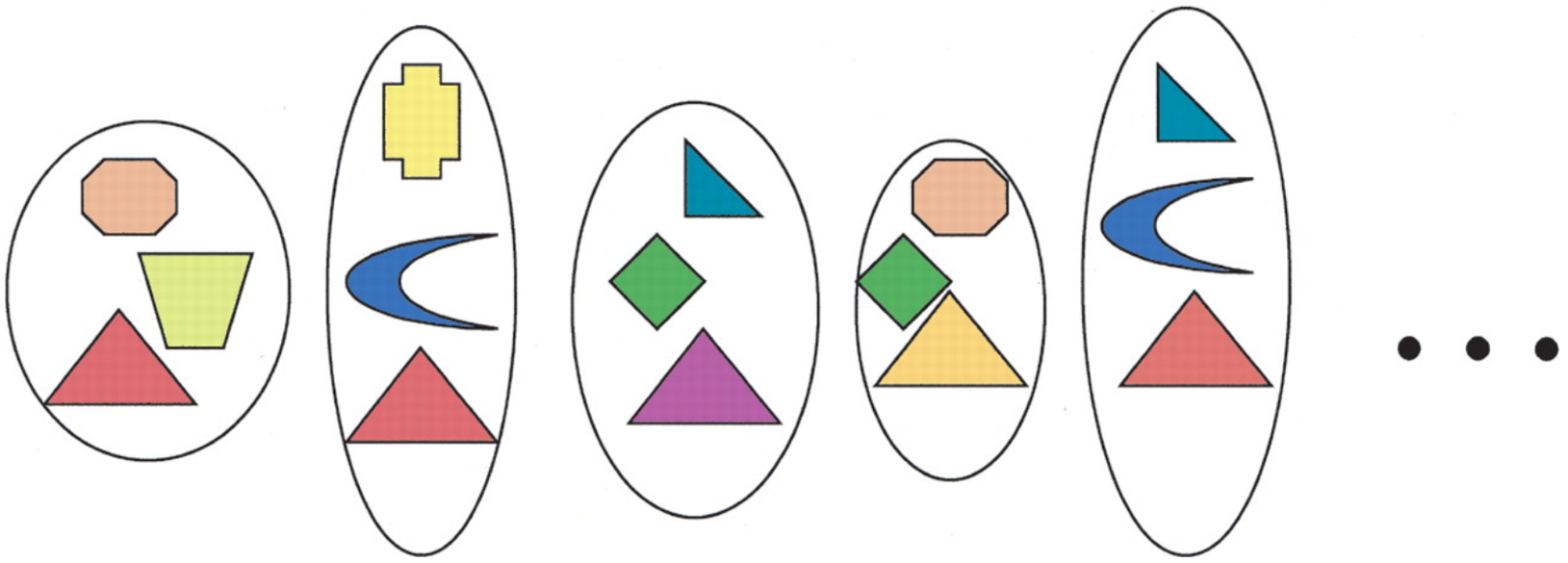
NOT NECESSARY



ENVELOPE

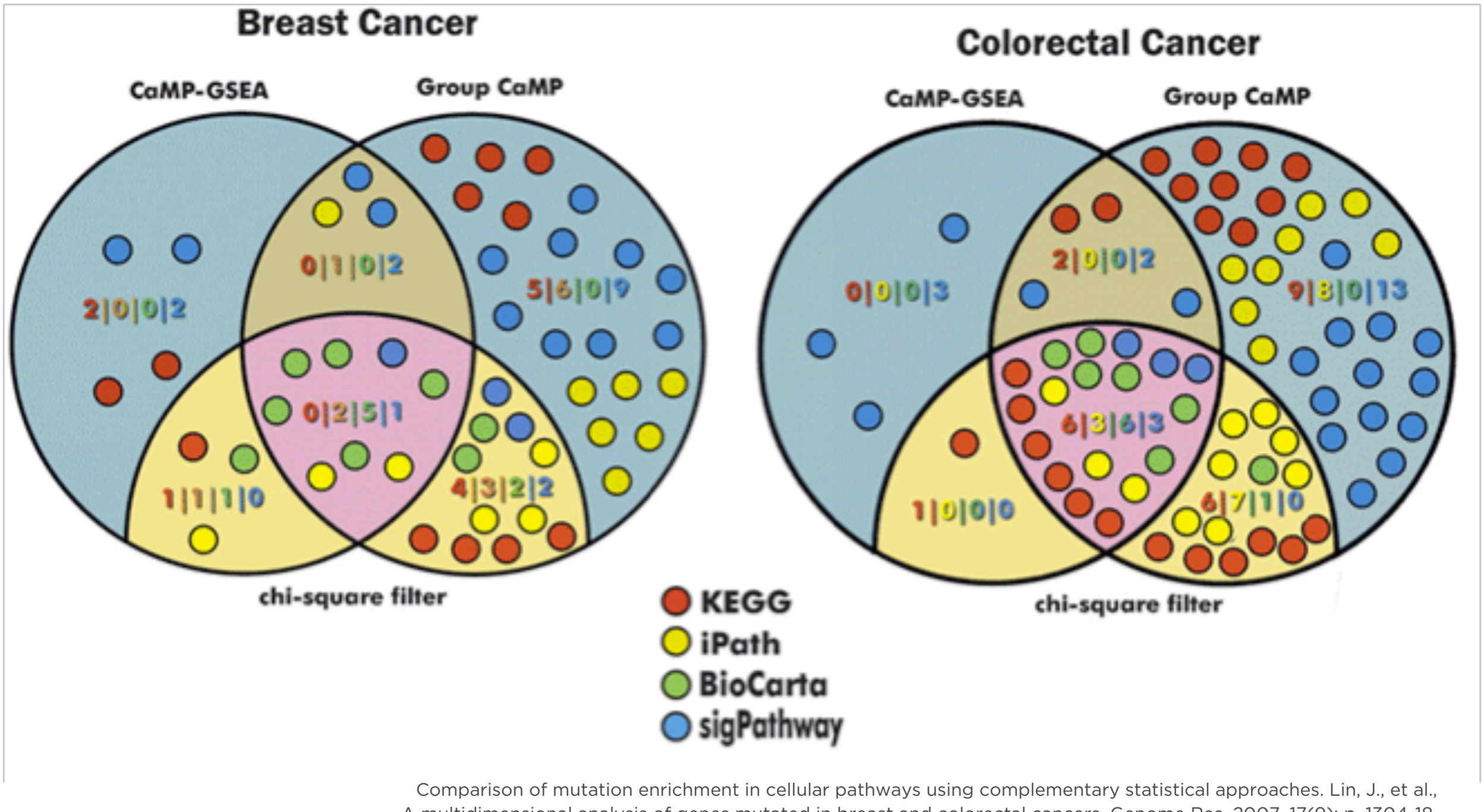
METABOLISM

TRANSLATION



A mix-and-match model for prokaryotic genome evolution. Charlebois, R.L. and W.F. Doolittle, Computing prokaryotic gene ubiquity: rescuing the core from extinction. Genome Res, 2004. 14(12): p. 2469-77.

NECESSARY, BUT NOT HELPFUL

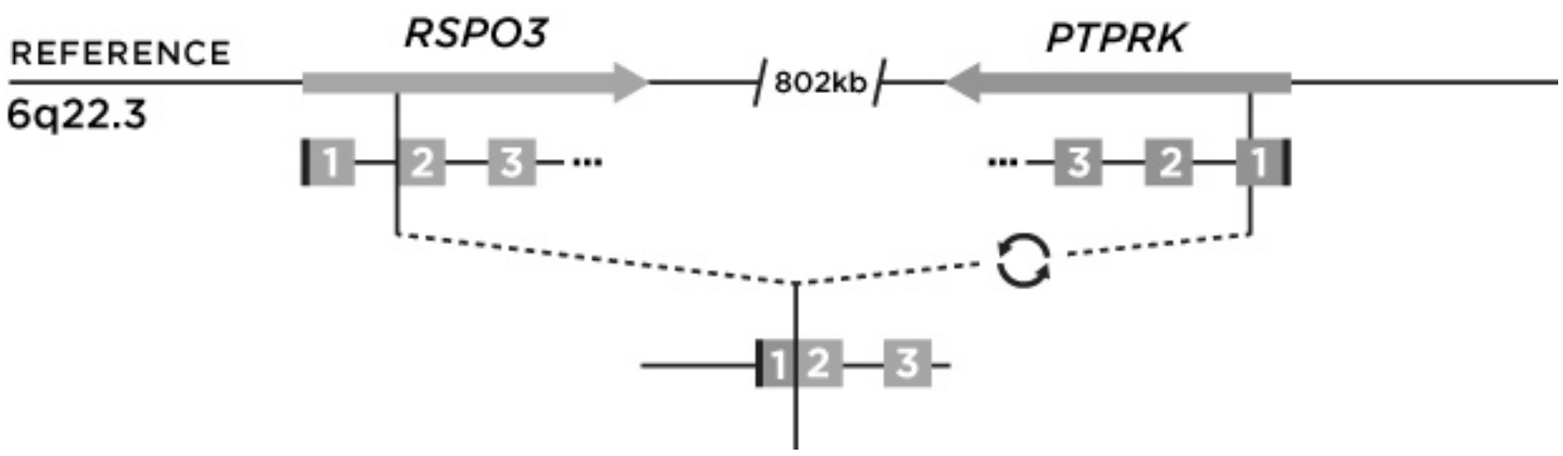


Comparison of mutation enrichment in cellular pathways using complementary statistical approaches. Lin, J., et al., A multidimensional analysis of genes mutated in breast and colorectal cancers. Genome Res, 2007. 17(9): p. 1304-18.

EXPLORATION VS COMMUNICATION

to explore data, use effective visual encodings

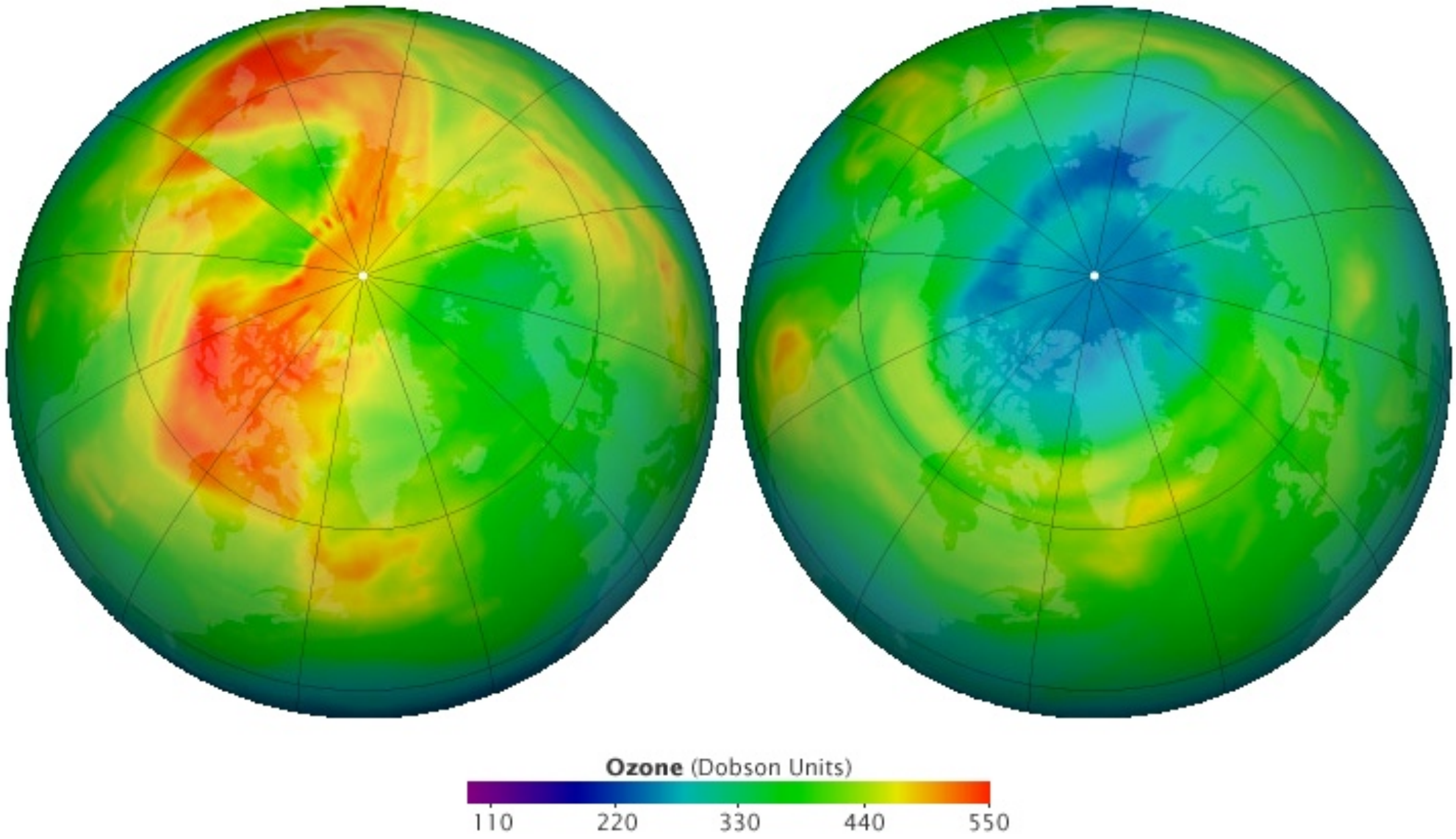
to communicate concepts, use effective design



visualization: communicating, clearly
MKWEB.BCGSC.CA

CONSEQUENCES OF BAD ENCODING

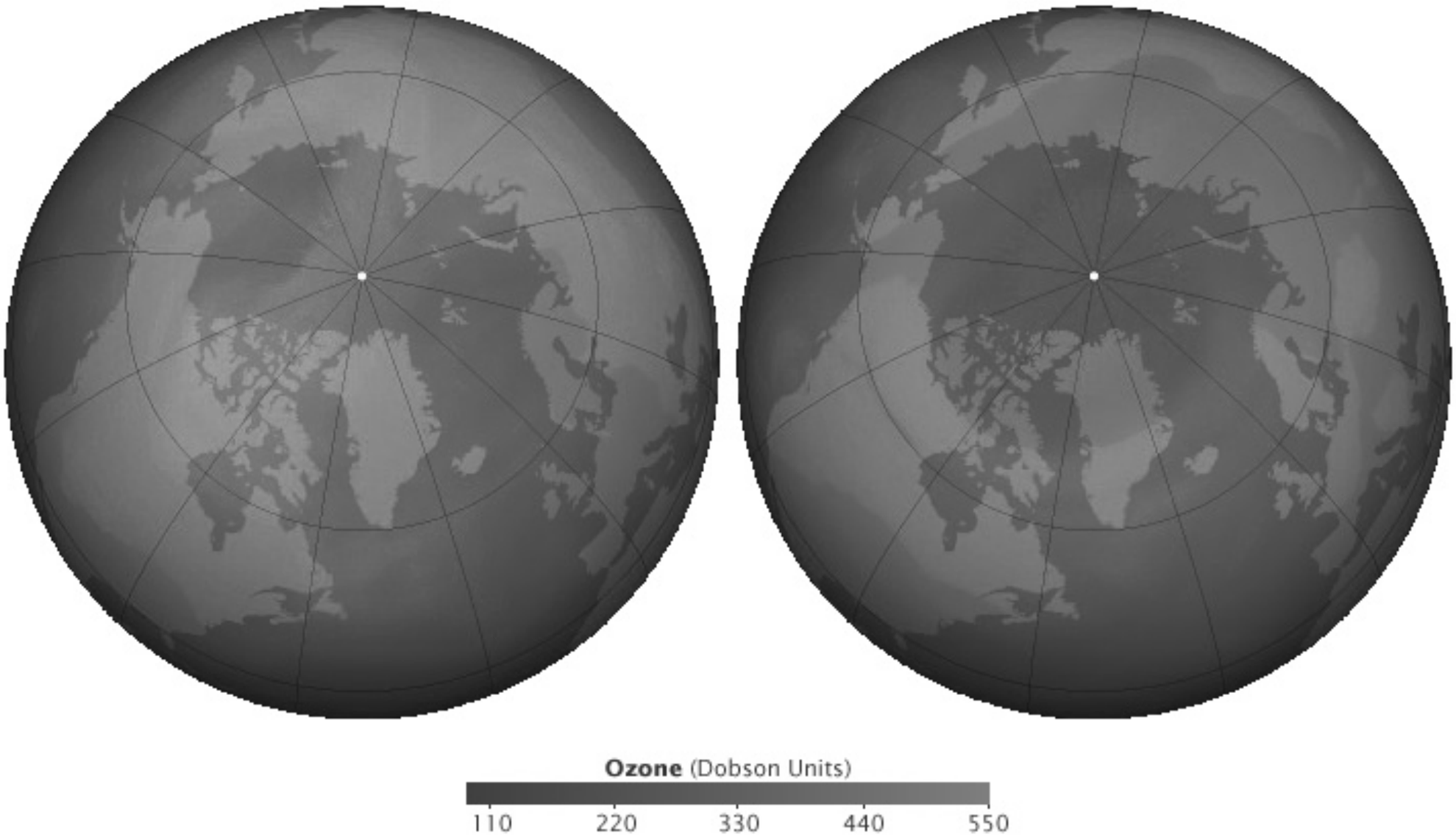
bad encoding doesn't mean the end of the world … maybe



Recent observations from satellites and ground stations suggest that atmospheric ozone levels for March in the Arctic were approaching the lowest levels in the modern instrumental era. *http://earthobservatory.nasa.gov/IOTD/view.php?id=49874*

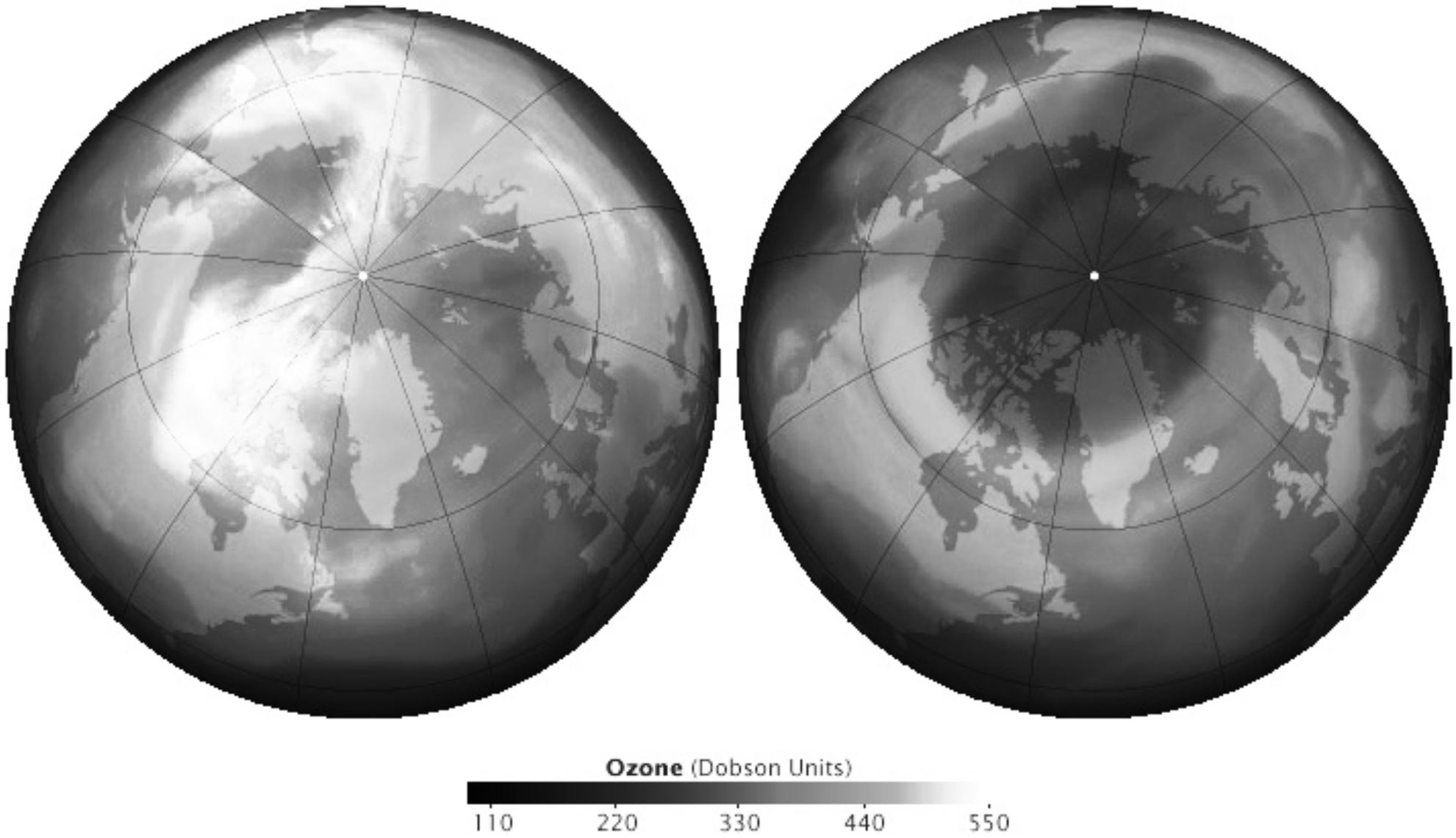
CONSEQUENCES OF BAD ENCODING

NYT did not use the figure – b/w conversion fail



CONSEQUENCES OF BAD ENCODING

use tone instead of hue

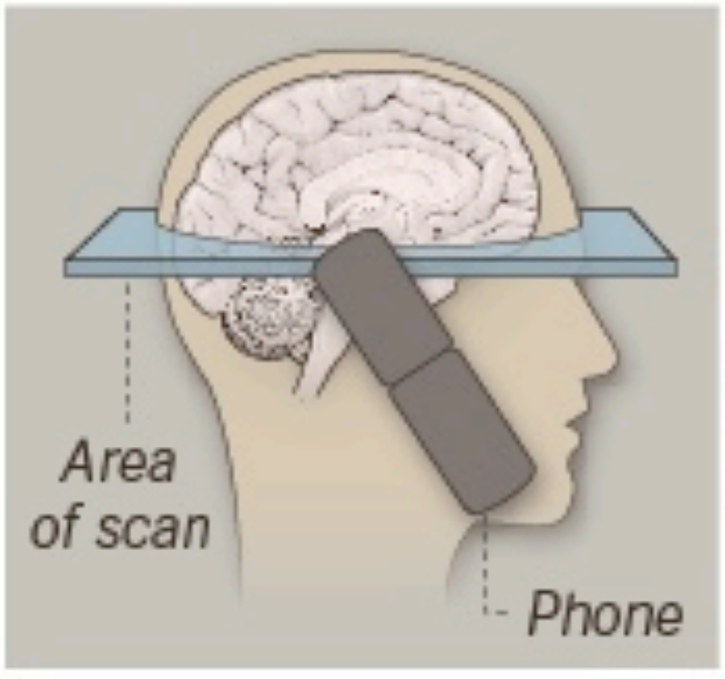


HEALTH IN THE HANDS OF A HEAT MAP

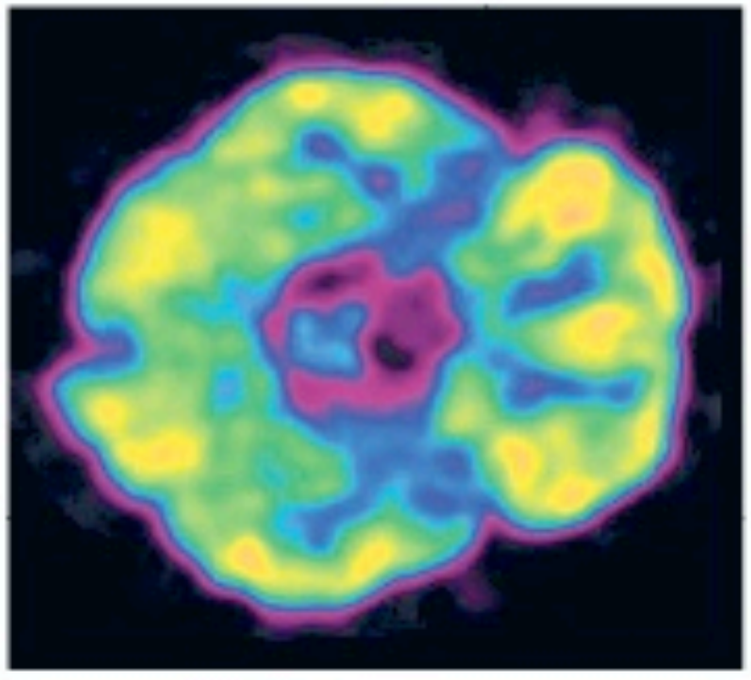
Cellphones and the Brain

Researchers tested 47 people by placing a cellphone at each ear. Both phones were off in one test, and in the other test the right phone was on a muted call. After 50 minutes, brain scans showed increased consumption of glucose, or sugar, in areas of the brain near the activated phone.

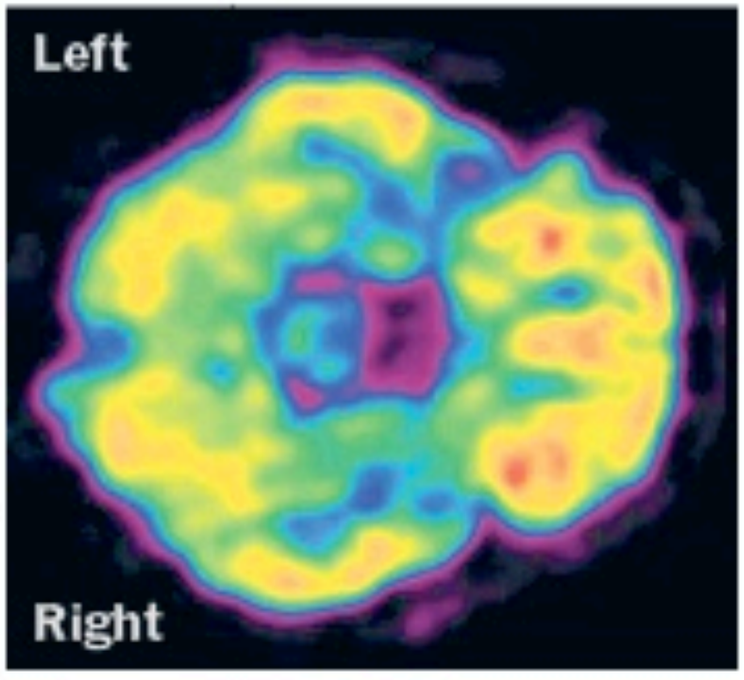
BRAIN SCAN



BOTH CELLPHONES OFF



RIGHT CELLPHONE ON



Rate of brain glucose metabolism LOW HIGH

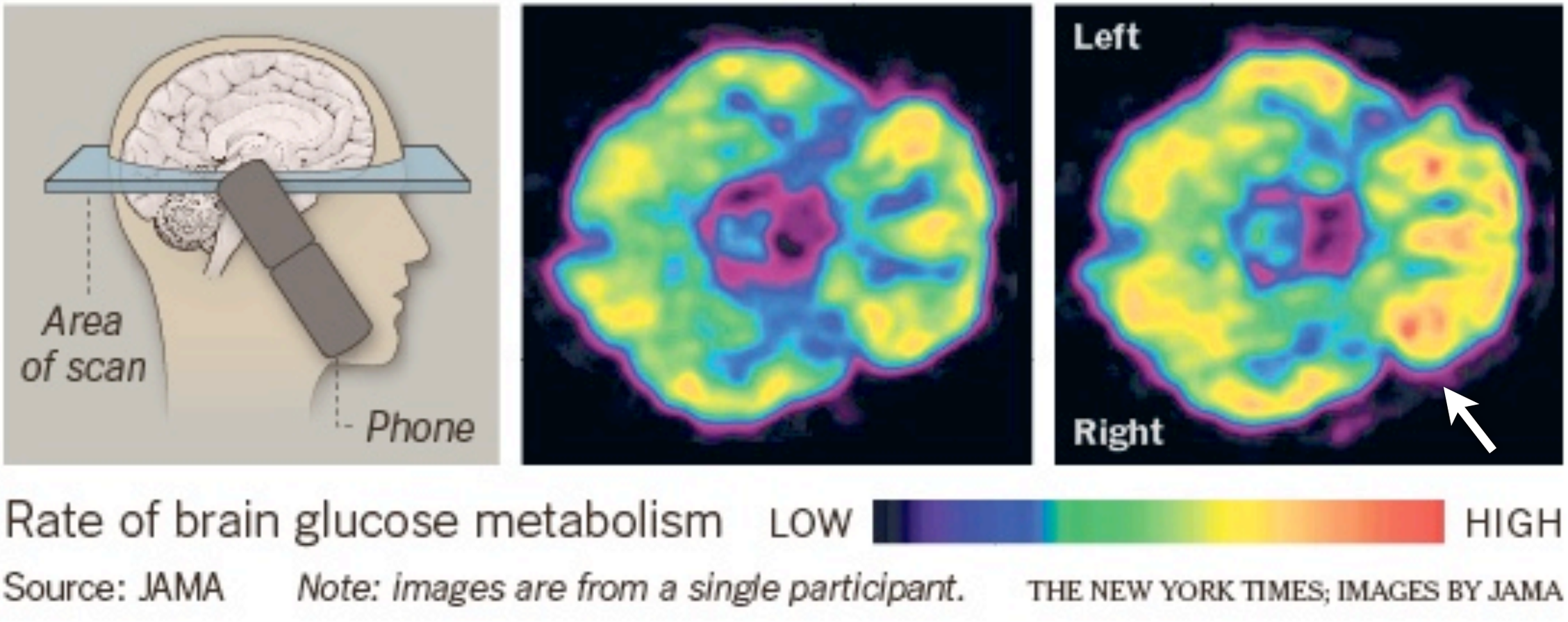
Source: JAMA *Note: images are from a single participant.* THE NEW YORK TIMES; IMAGES BY JAMA

http://well.blogs.nytimes.com/2011/02/22/cellphone-use-tied-to-changes-in-brain-activity/

HEALTH IN THE HANDS OF A HEAT MAP

Cellphones and the Brain

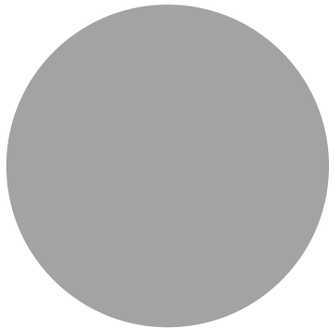
Researchers tested 47 people by placing a cellphone at each ear. Both phones were off in one test, and in the other test the right phone was on a muted call. After 50 minutes, brain scans showed increased consumption of glucose, or sugar, in areas of the brain near the activated phone.

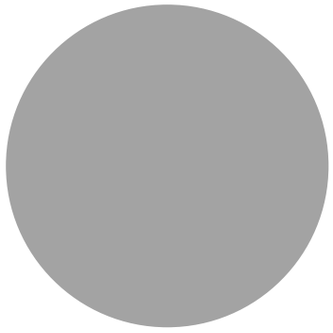


http://well.blogs.nytimes.com/2011/02/22/cellphone-use-tied-to-changes-in-brain-activity/

IT'S A FEATURE, NOT A BUG

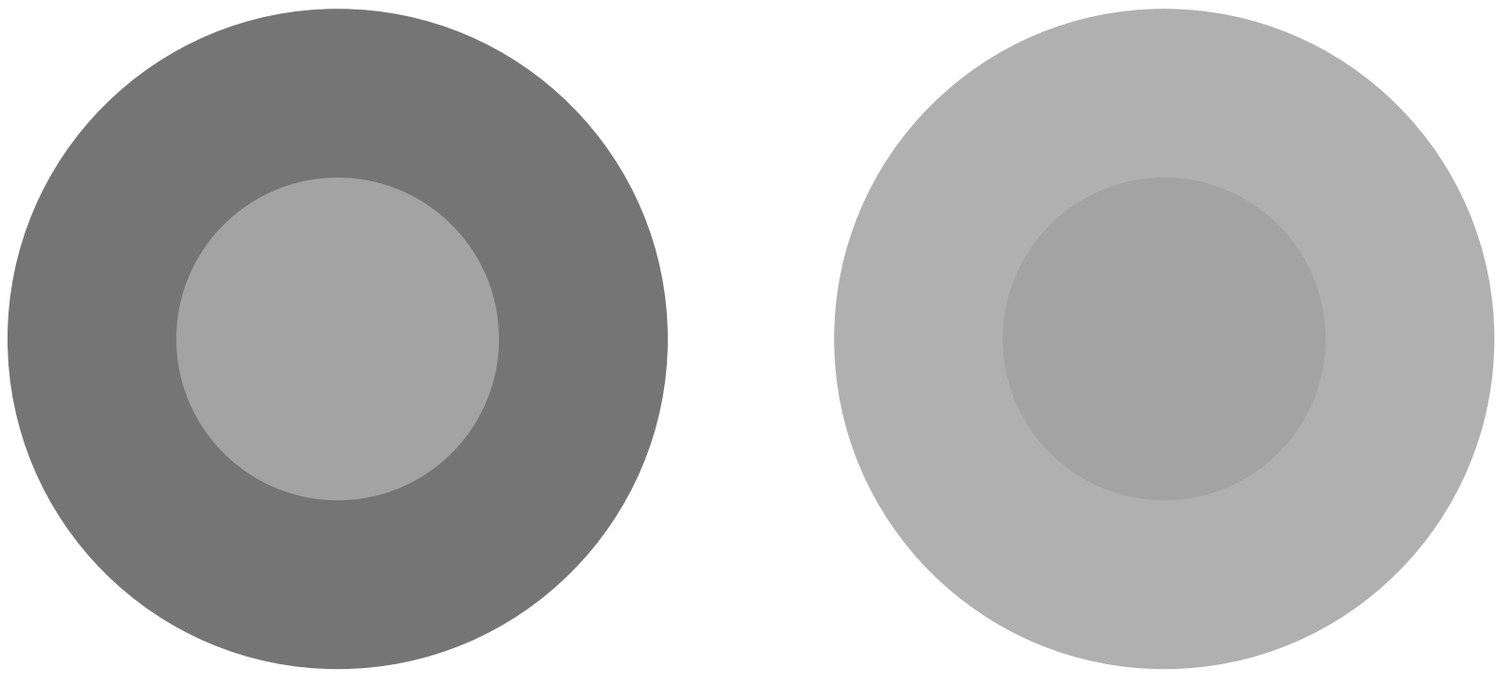
visual system preprocessing ...





IT'S A FEATURE, NOT A BUG

can powerfully impact perception causing us to unknowingly commit errors

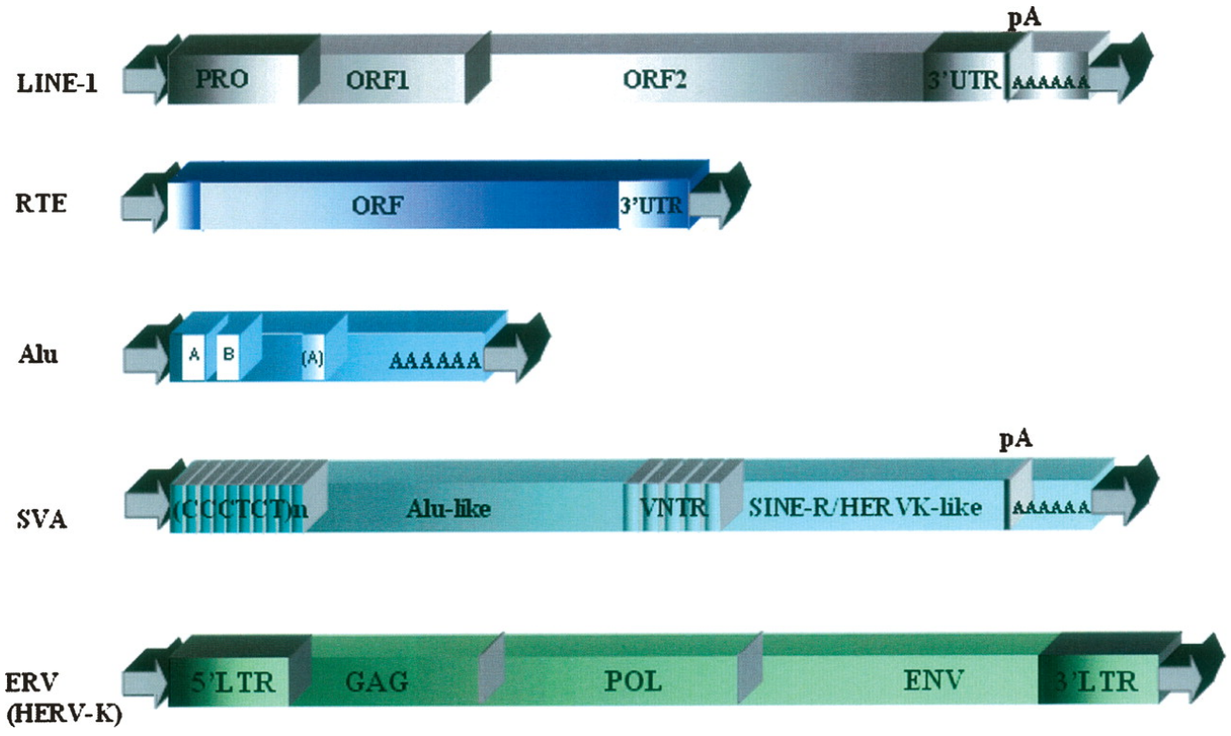


This effect has significant consequences in how we perceive and interpret heat maps, which require us to judge the relative tone of neighbouring colors. Points of view: Heat maps. Nils Gehlenborg & Bang Wong, *Nature Methods* 9, 213 (2012).

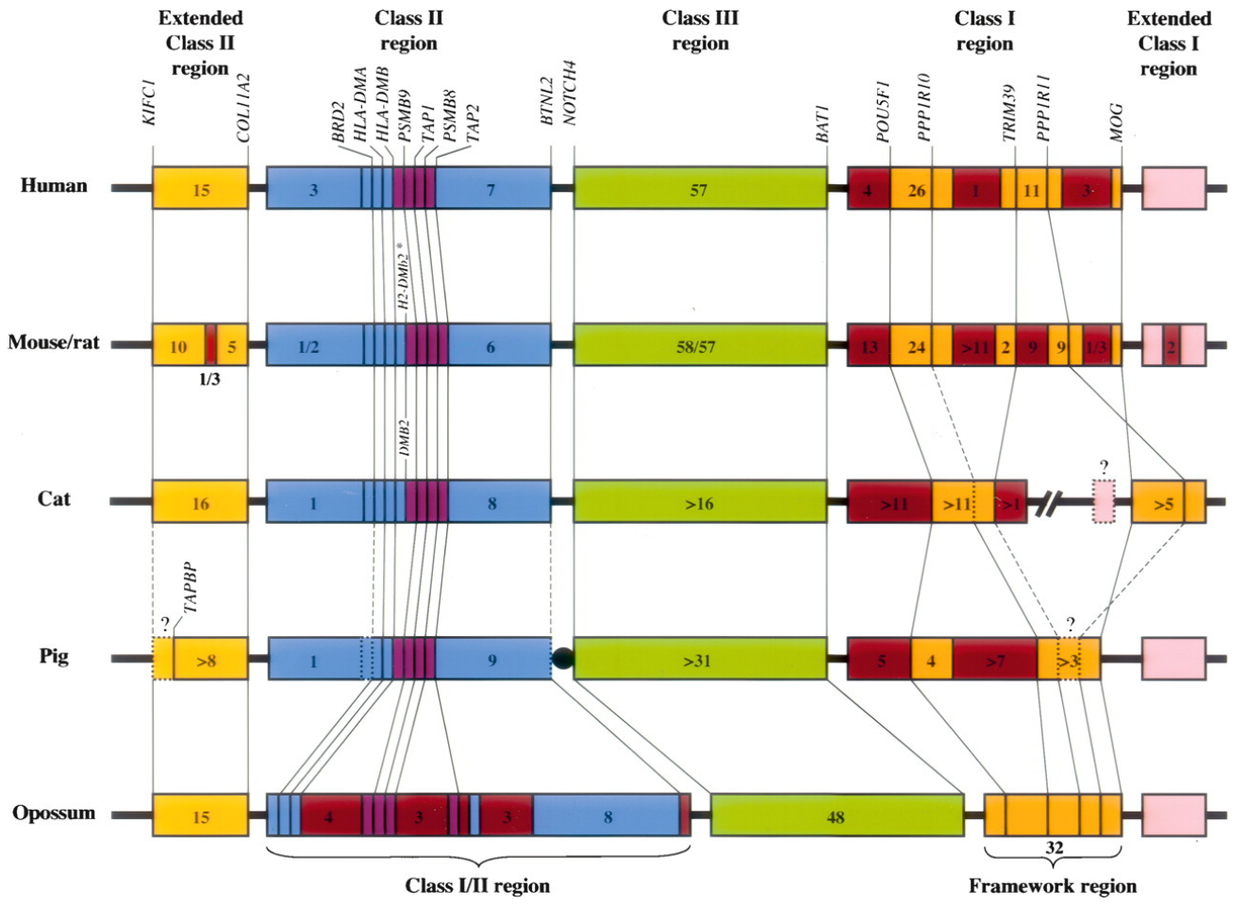
visualization: communicating, clearly
MKWEB.BCGSC.CA

OBJECTIVE ASPECTS OF ATTRACTION

OBJECTIVE ASPECTS OF CLEAR COMMUNICATION



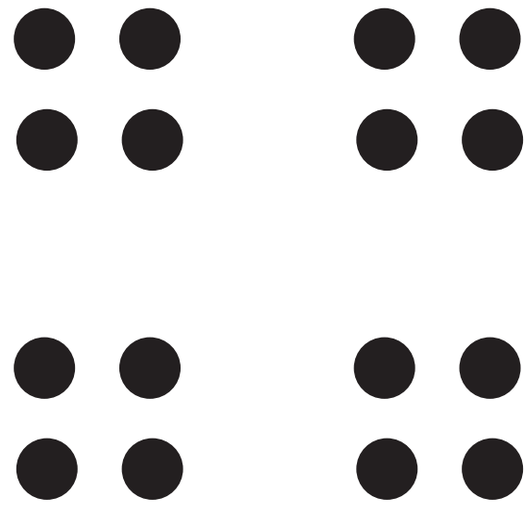
Gentles, A.J., et al., Evolutionary dynamics of transposable elements in the short-tailed opossum Monodelphis domestica. Genome Res, 2007. 17(7): p. 992-1004.



Samollow, P.B., The opossum genome: insights and opportunities from an alternative mammal. Genome Res, 2008. 18(8): p. 1199-215.

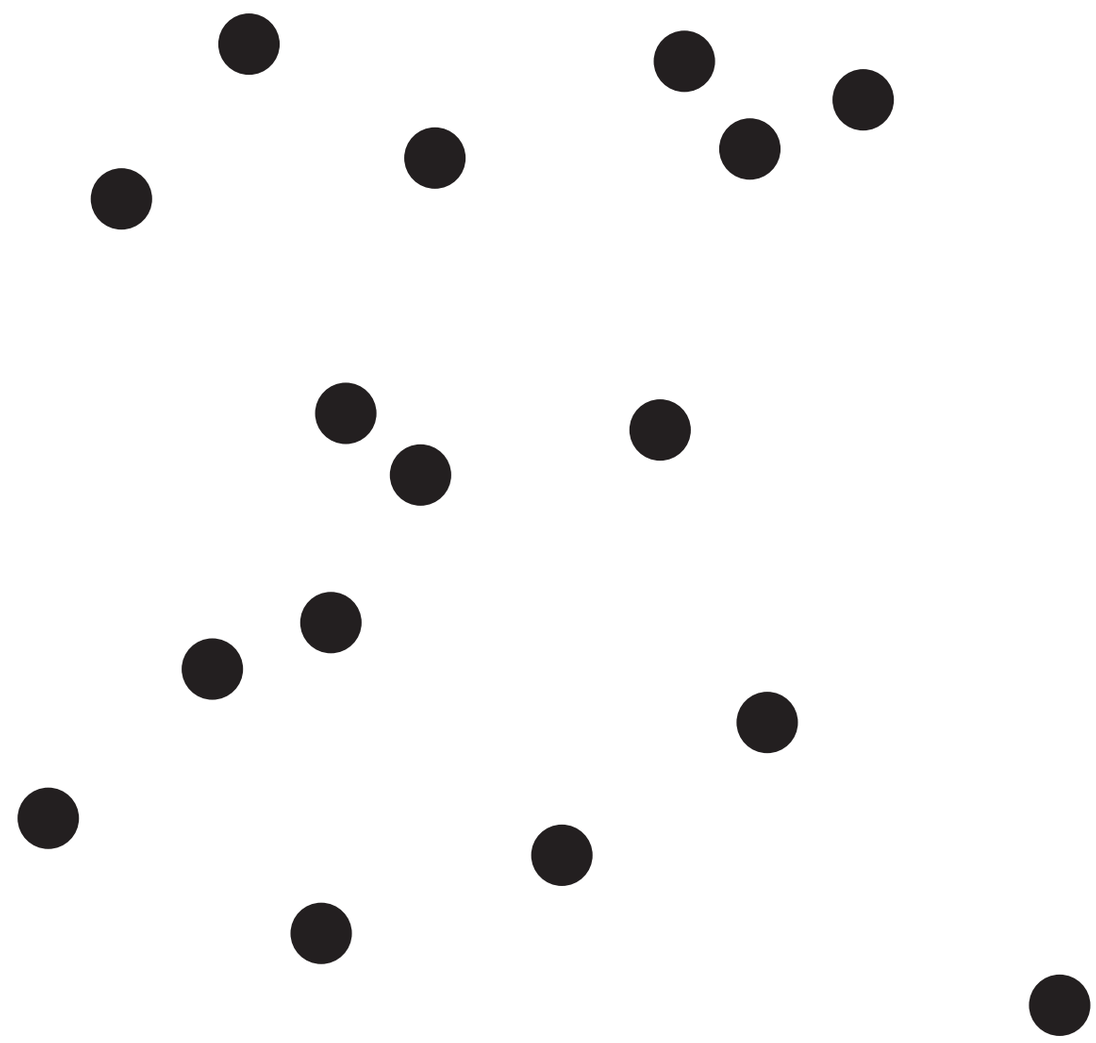
When creating figures, there are certain rules that we can follow that enhance the clarity of the message. Elements should always be legible and consistently formatted. Repetition should be limited, as should unnecessary variation.

REMOVE CLUTTER



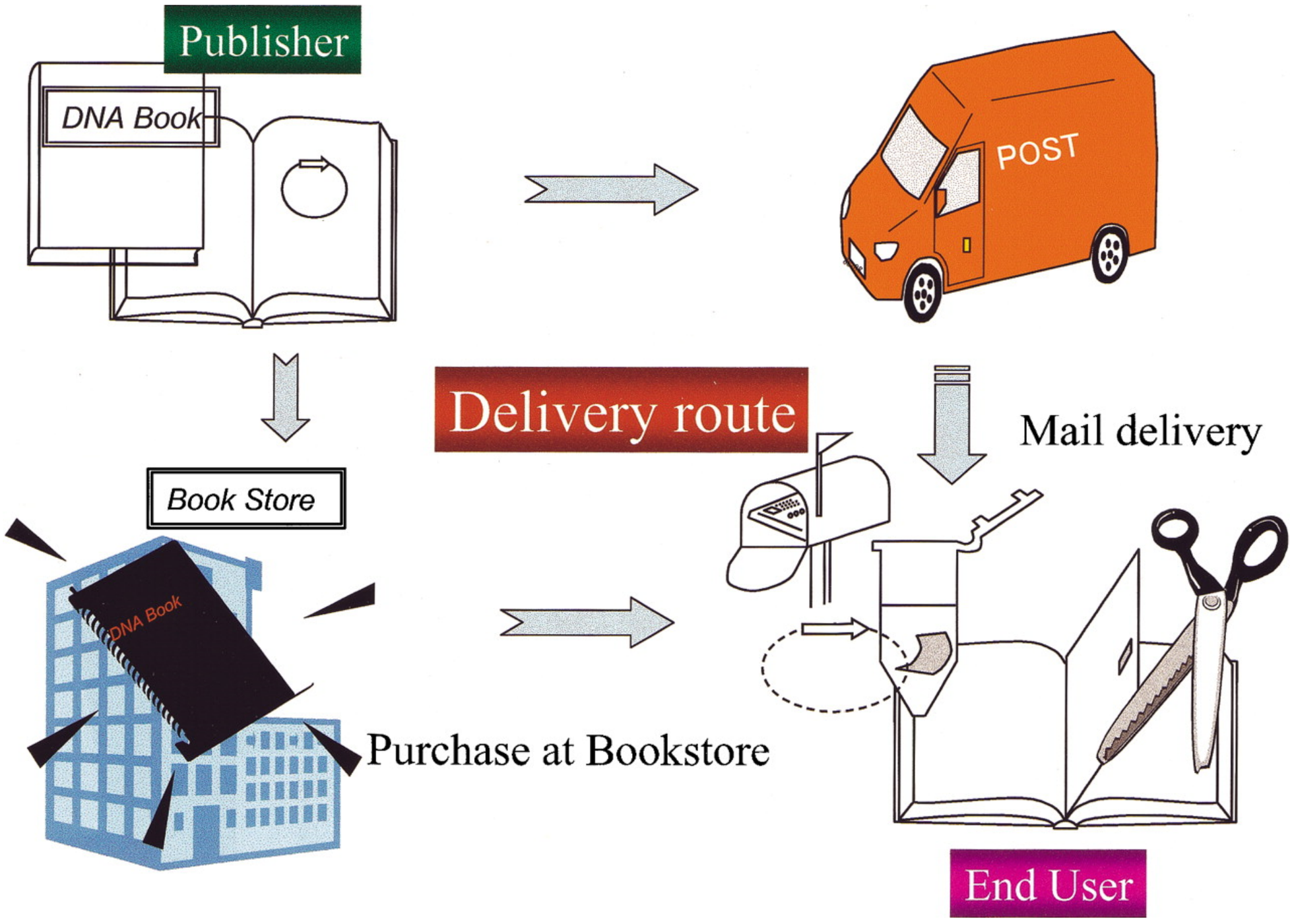
How long did it take you to count the dots? Did you notice that you did not have to explicitly enumerate them? Objects clustered into small groups can be processed pre-attentively, allowing you to “count” the dots quickly. Even more importantly, your confidence that you counted correctly is high.

REMOVE CLUTTER



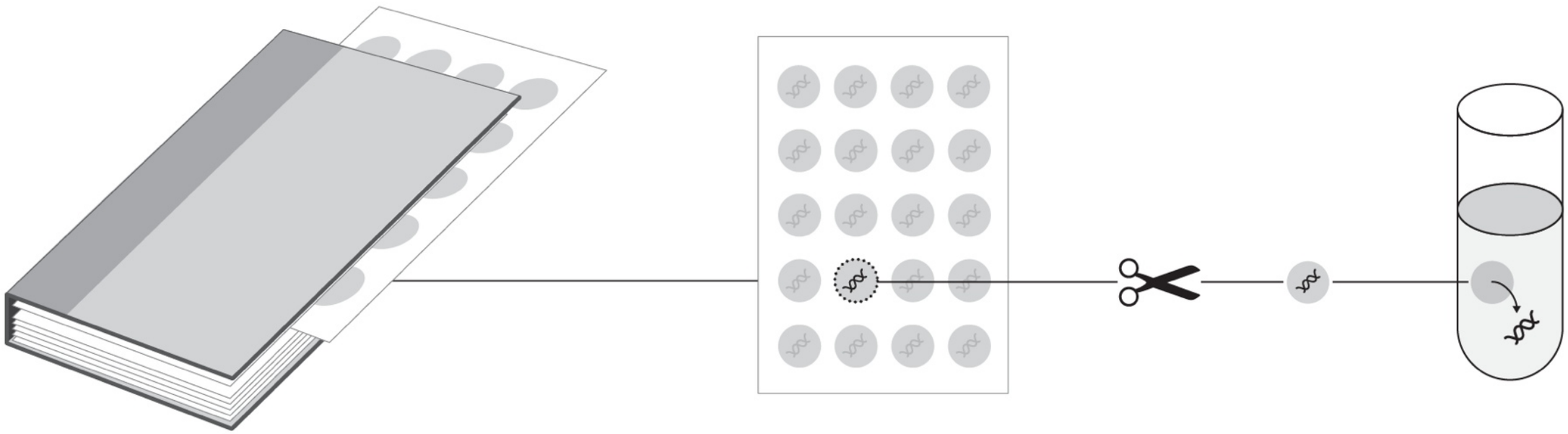
Counting the dots here takes an order of magnitude longer. And when finished counting, your confidence in having a correct count is lower than in the previous example. This example explicitly demonstrates the benefit of visually organizing information: faster parsing with greater confidence.

FOCUS ON THE CORE MESSAGE



Concept of the “DNA Book.” Kawai, J. and Y. Hayashizaki, DNA book. Genome Res, 2003. 13(6B): p. 1488-95.

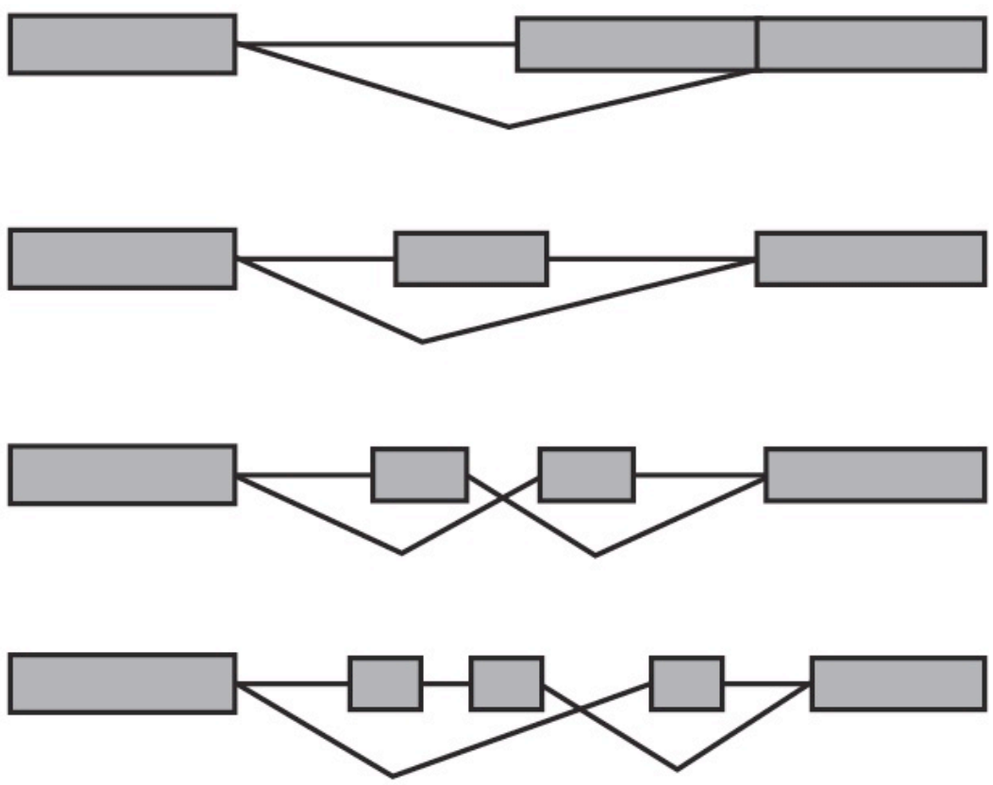
FOCUS ON THE CORE MESSAGE



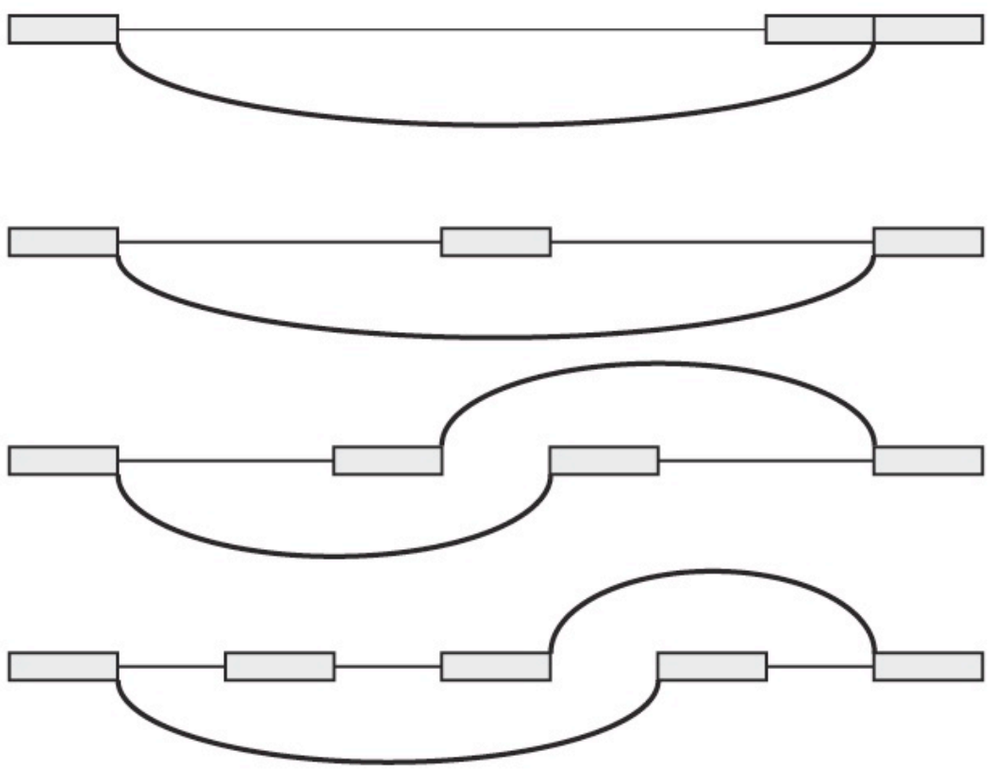
The concept of the DNA book is minimally represented in this figure. Core ideas are shown with familiar metaphors.

REMOVE UNNECESSARY DEGREES OF FREEDOM

spacing variation is implied



variation refactored

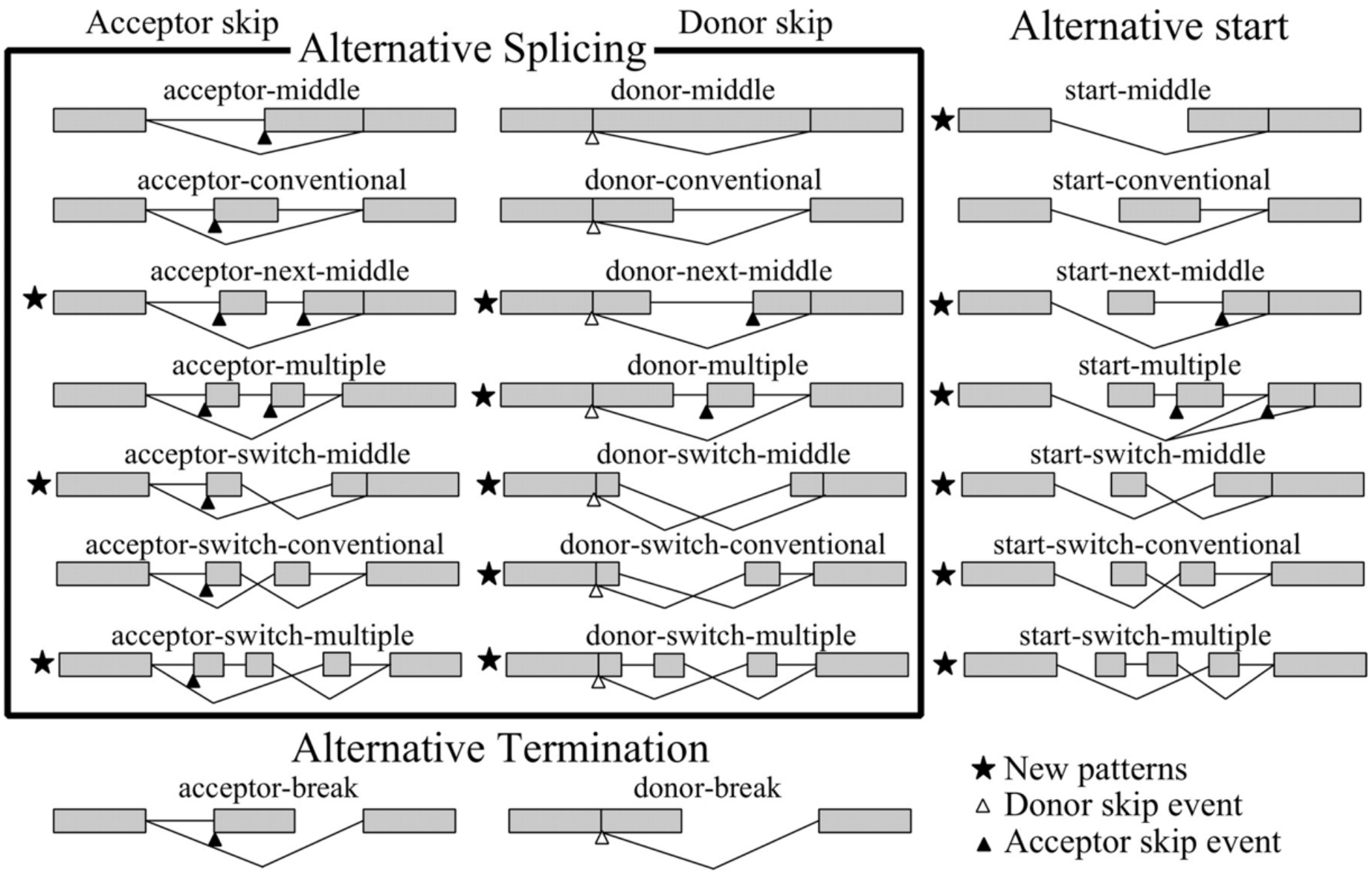


Sharov AA, Dudekula DB, Ko MS (2005) Genome-wide assembly and analysis of alternative transcripts in mouse. Genome Res 15: 748-754.

The reader does not know what is important - it is up to the author to indicate it. When presented with variation in the figure the reader first assumes that it encodes important information. When this assumption is invalidated, the reader's focus is detracted.

In this example of splicing, exon size and spacing is unimportant. Remember, splicing is a statement about adjacency, not about the size and distribution of exons.

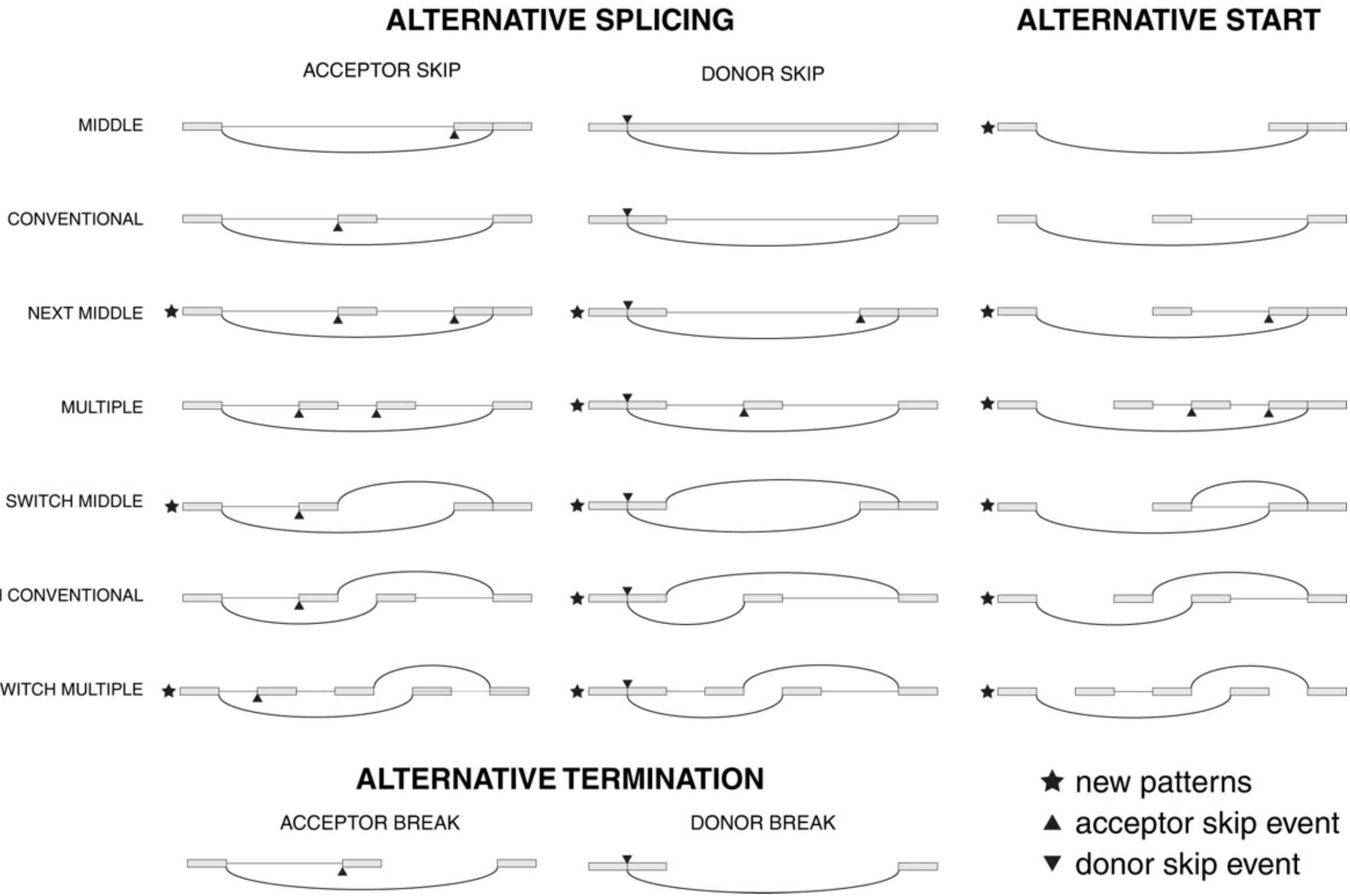
RESPECT NATURAL HIERARCHIES



Sharov AA, Dudekula DB, Ko MS (2005) Genome-wide assembly and analysis of alternative transcripts in mouse. Genome Res 15: 748-754.

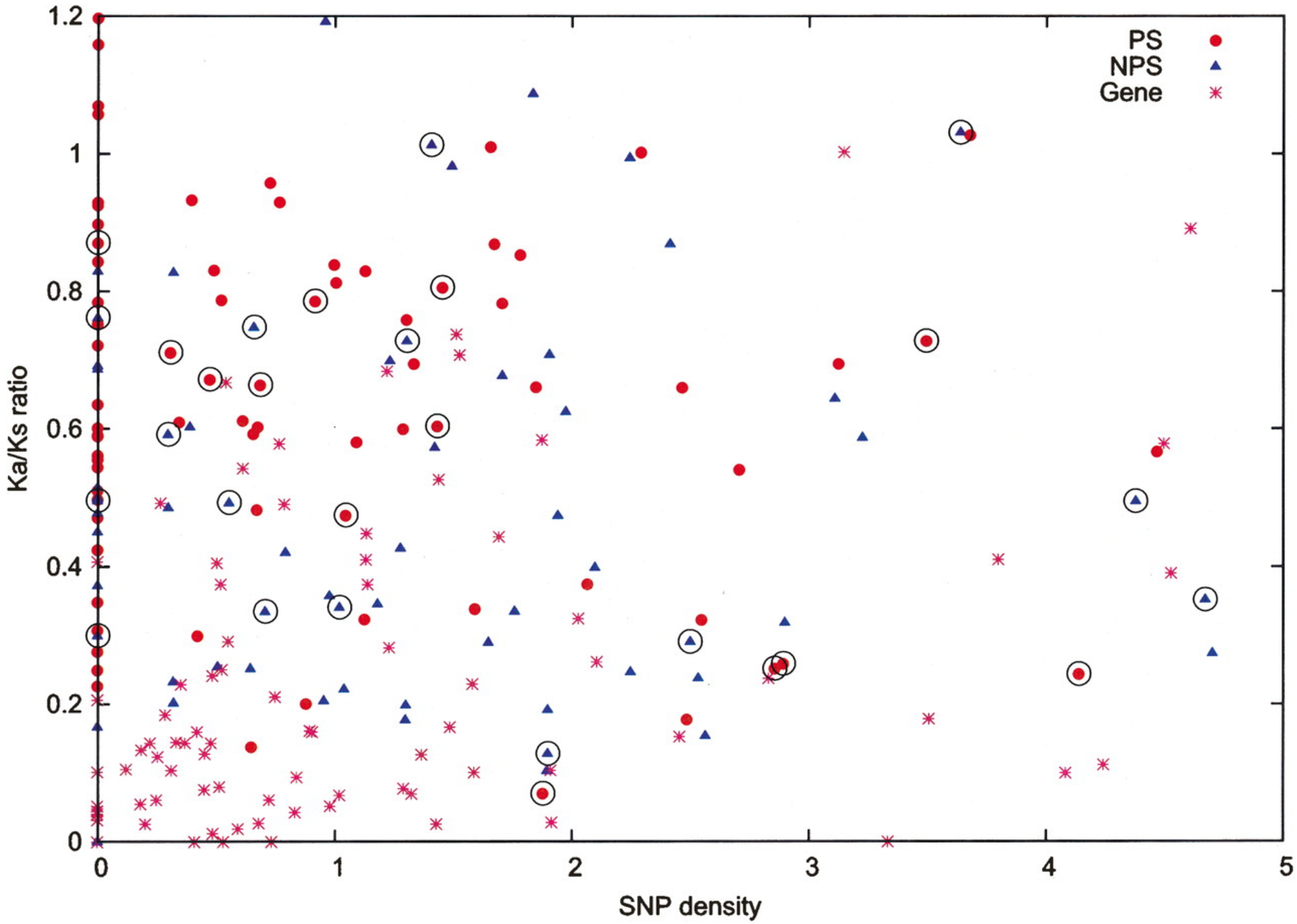
When information can be presented in a table, do so. Here, labels are repeated and their pattern is difficult to discern. The reader must expend significant energy in transforming the figure into an organized mind map. This task is made very difficult by clutter: unnecessary variation in visual elements and redundancy in labels.

RESPECT NATURAL HIERARCHIES



visualization: communicating, clearly
MKWEB.BCGSC.CA

RESPECT NATURAL HIERARCHIES



Comparison of Ka/Ks ratio and SNP density for genes and pseudogenes. Zheng, D., et al., Pseudogenes in the ENCODE regions: consensus annotation, analysis of transcription, and evolution. Genome Res, 2007. 17(6): p. 839-51.

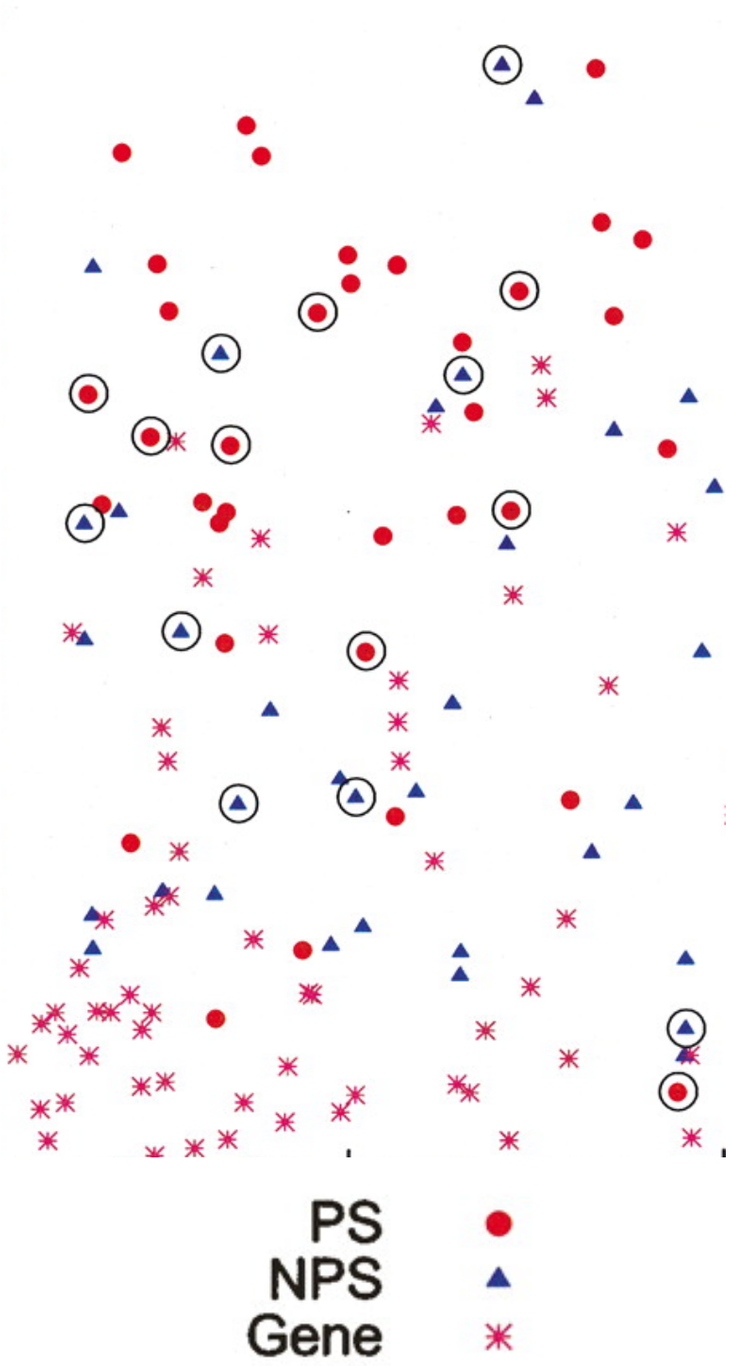
In the same way that hierarchical information should be tabularized, where necessary, the choice of visual encoding should reflect the hierarchy in a way that allows for easy visual access to each level of information. In this example, it is difficult to maintain focus on the circled red circles (transcribed processed genes) - the circled triangles compete for attention. The circular outline, which does not appear in the legend, has high salience and inhibits perception of variation in gene type.

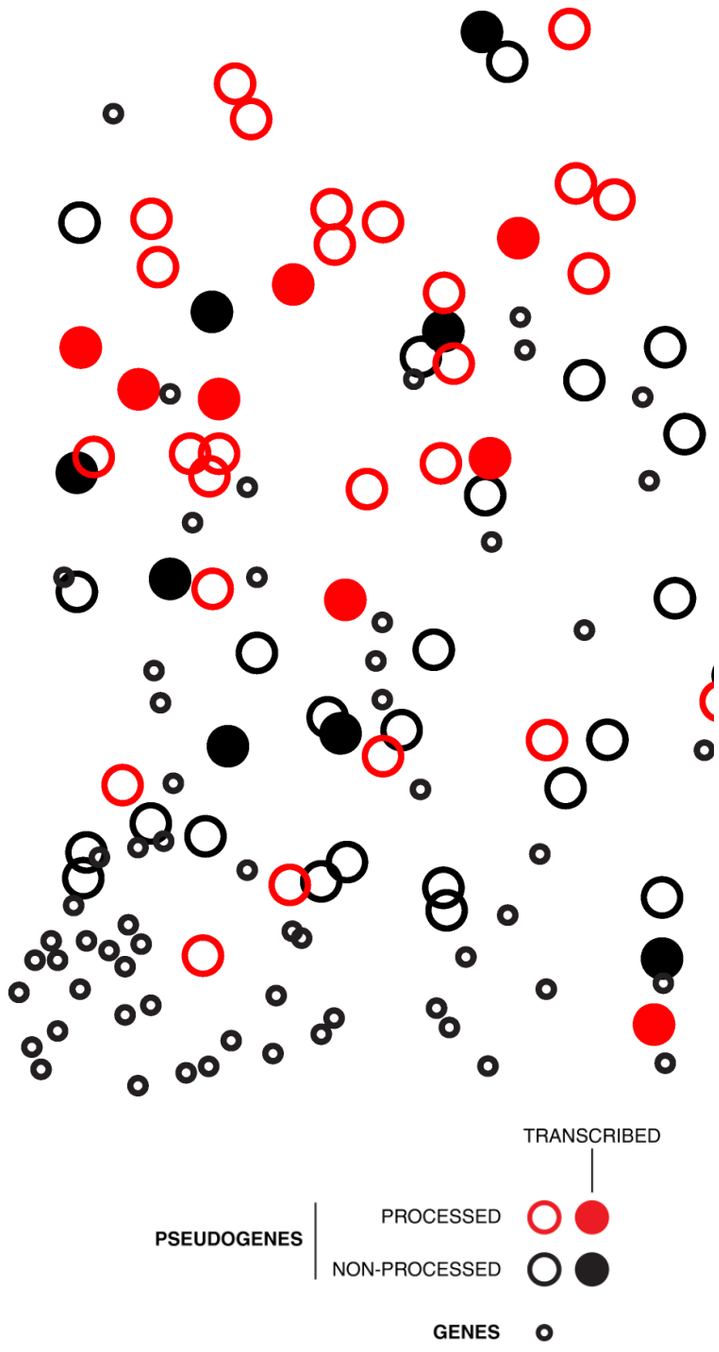
RESPECT NATURAL HIERARCHIES

The original key is not intuitive and ambiguous. Is a red star less or more important than a blue triangle? The redesigned encoding is a better mapping between salience (what we see first) with relevance (what is most important). The encoding uses shape and color that is intuitive: small/large, hollow/solid, black/red has less/more visual weight.

visualization: communicating, clearly
MKWEB.BCGSC.CA

RESPECT NATURAL HIERARCHIES





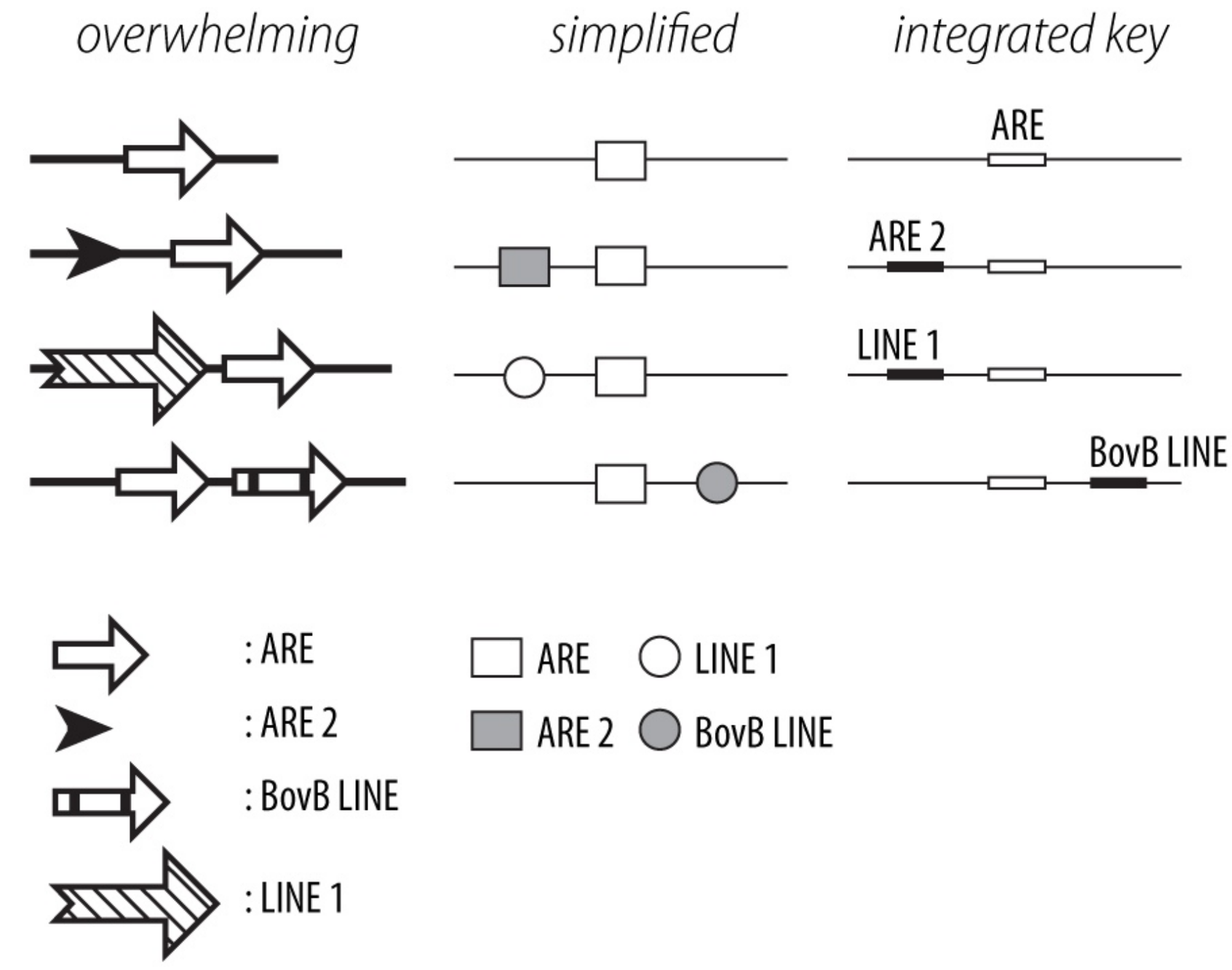
The new encoding makes it possible to maintain focus on any of the gene groups. We can separate the figure into information channels without disrupting cross-talk.

Try focusing on the following groups of genes:

- processed/transcribed (solid red)
- non-processed (hollow black)
- uncategorized (small hollow black)

This redesign expands on the previous example of counting 16 dots. Here, you can locate all the genes in a group quickly and do so with confidence.

REMOVE REDUNDANCY TO REVEAL PATTERNS



You don't need arrows if they all point in the same direction.

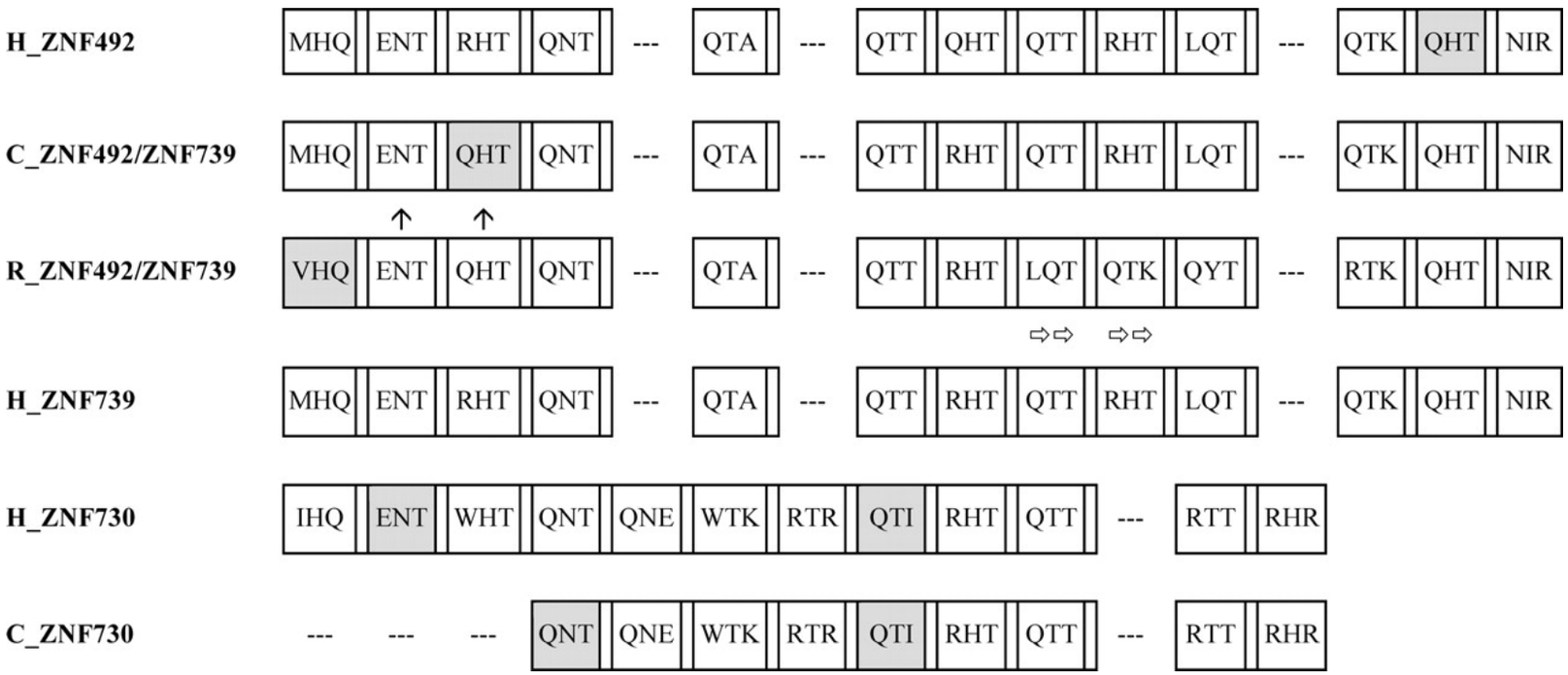
There is a lot of variation in position and size of elements which distracts from identifying patterns. The paper does not include an explicit comparison of the size of ARE and LINE elements, which makes the arrow size in the figure ambiguous. Is arrow size encoding the size of the element?

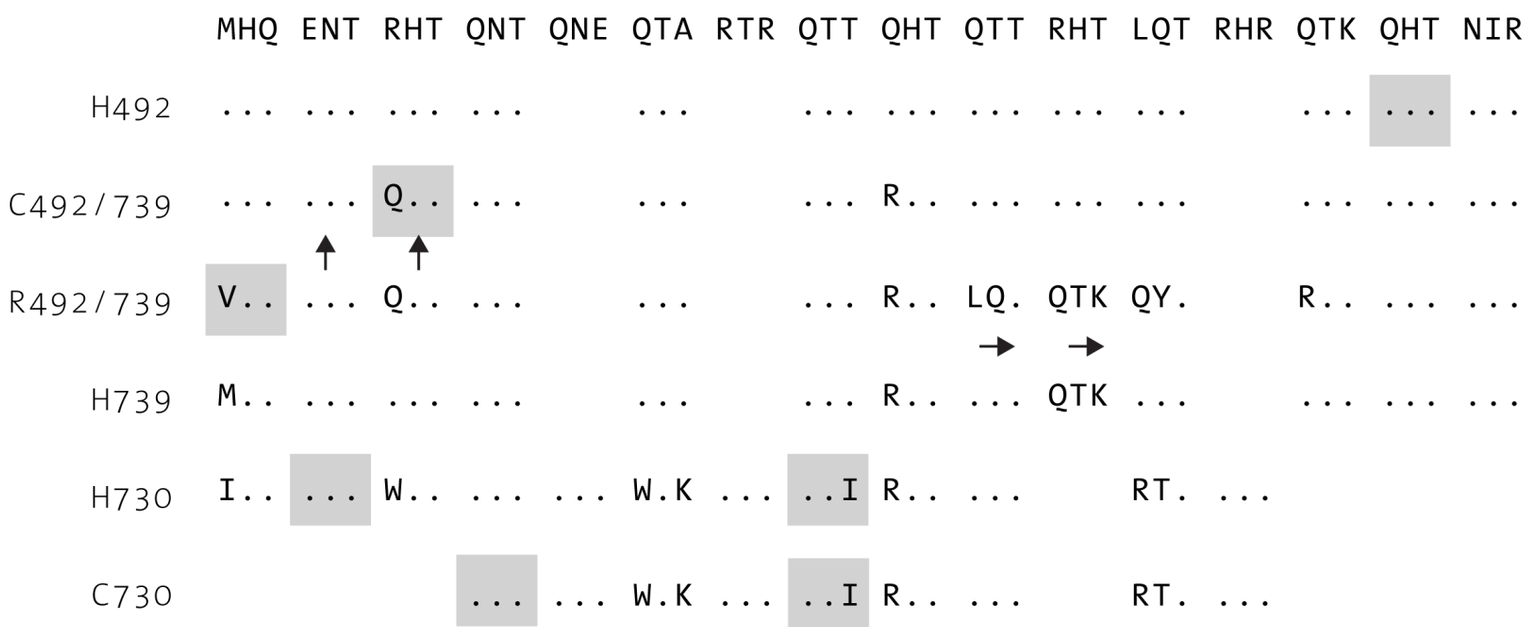
When elements in the figure appear only once, consider integrating the key into the figure to make lookup faster. For example, despite that ARE 2 appears only once, in the "simplified" version the ARE 2 glyph appears twice (once in the figure and once in the legend).

The "integrated key" version uses the fact that labels are sufficient to distinguish the non-ARE elements - changing their glyphs (or line color) is unnecessary. Because the ARE element repeats for each case, it is made hollow to increase the emphasis on the other elements.

Nikaido M, Rooney AP, Okada N (1999) Phylogenetic relationships among cetartiodactyls based on insertions of short and long interspersed elements: Hippopotamuses are the closest extant relatives of whales. Proceedings of the National Academy of Sciences 96: 10261-10266.

REMOVE REDUNDANCY TO REVEAL PATTERNS





Zinc finger exon analysis for ZNF493 and ZNF738, two divergent genes from the ZNF431 clade. Hamilton, A.T., et al., Evolutionary expansion and divergence in the ZNF91 subfamily of primate-specific zinc finger genes. Genome Res, 2006. 16(5): p. 584-94.

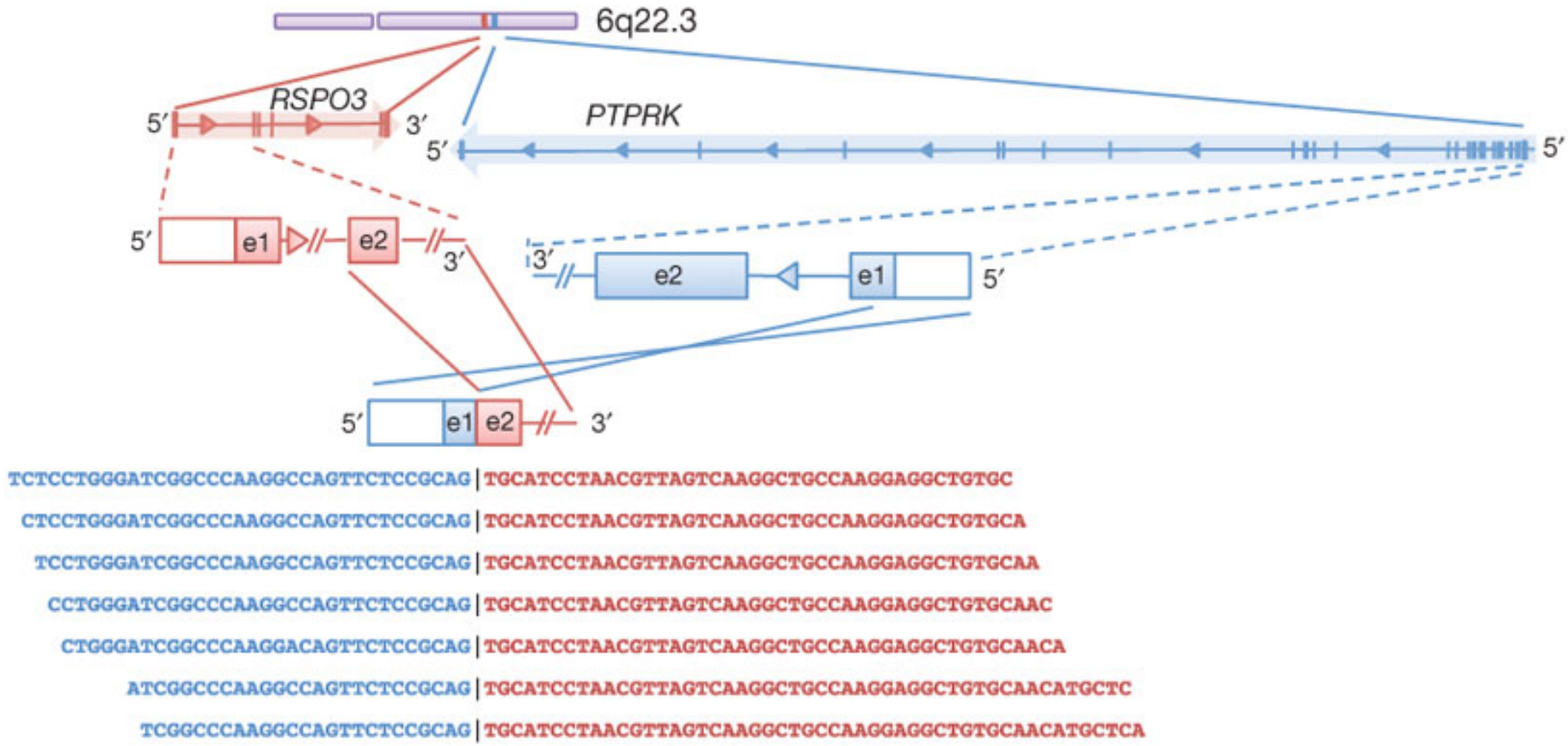
It is very difficult to identify small differences in complex encodings, particularly text.

Effective visualizations emphasize important differences and minimize the visual weight of redundant information. This figure contains both unnecessary redundancy (identical sequence is repeated, hiding differences) and unnecessary variation (different arrows) It also suffers from excess of organizational elements, such as boxes bounding each triplet and "...".

By using negative space to group elements and by factoring out consensus sequence, the presentation becomes more organized and parsable. Just as with the 16 dot example, the reader's confidence in identifying patterns will be decreased by the clutter of the original version.

The utility of a visualization is proportional to the number of questions it can answer, and the level of speed and confidence at which it can be done. Simple questions such as "Which triplet is changed in the most proteins relative to H492?" is easily answerable in the redesign (first QHT -> RHT in all proteins), but essentially impossible to visually assess in the original.

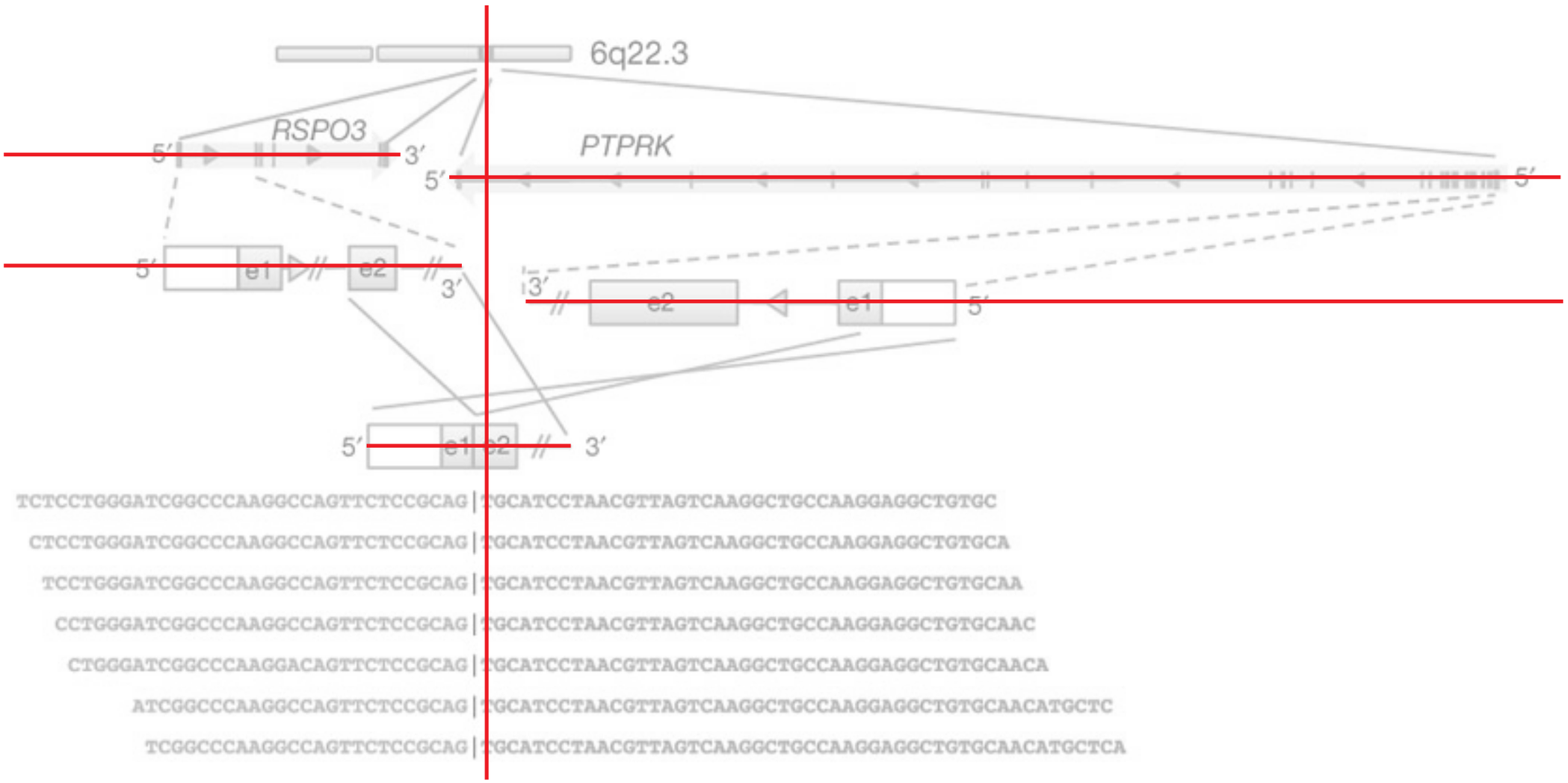
APPLYING THE PRINCIPLES



Recurrent R-spondin fusions in colon cancer. Somasekar Seshagiri *et al. Nature* 488, 660–664

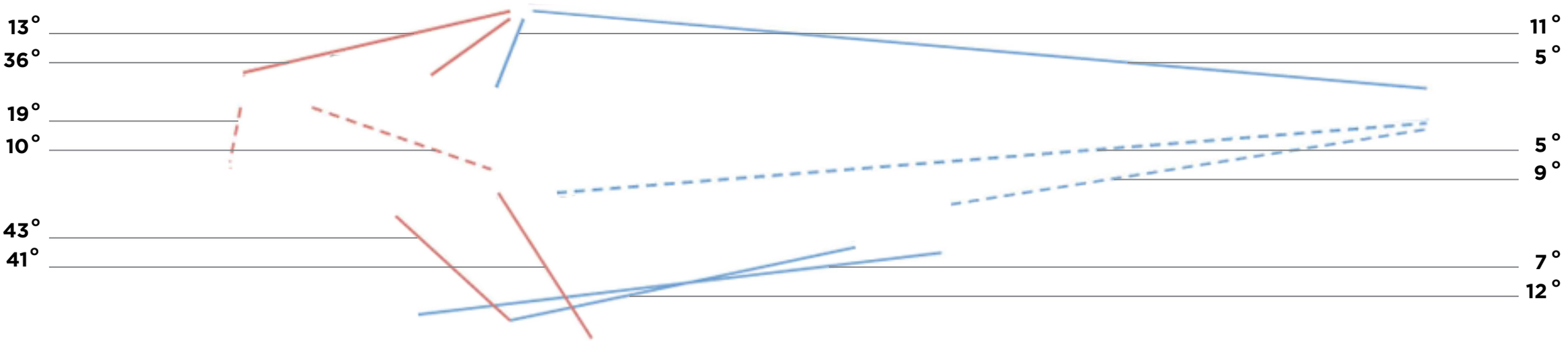
I will use this figure to illustrate how the principles exemplified in the previous slides can be used to improve figures.

DEGREES OF FREEDOM — SPATIAL



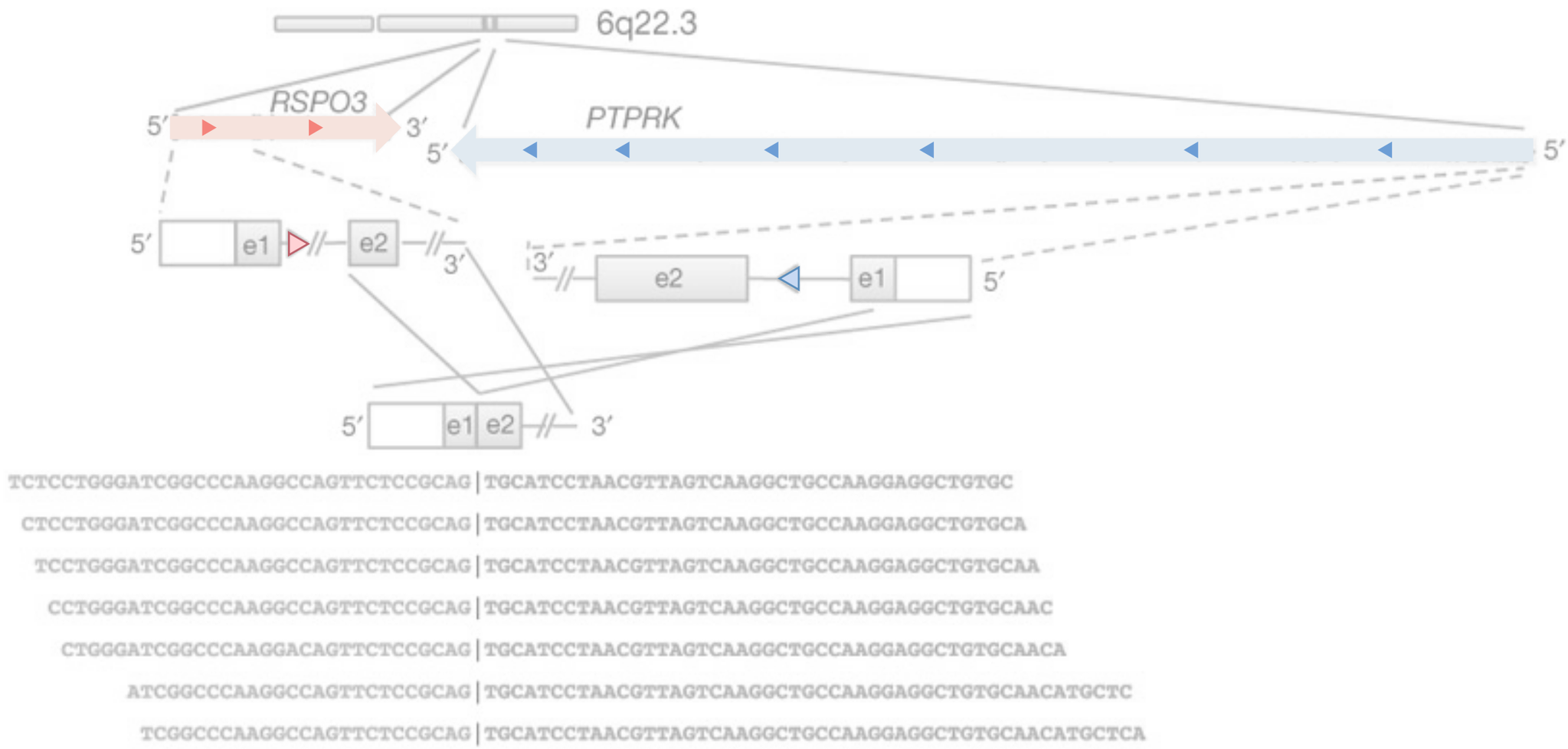
Maintaining good visual organization is important in helping the reader assess the hierarchy of elements as well as to emphasize positions of equivalent priority across multiple objects. For example the two genes RSPO3 and PTPRK are found at different vertical positions, which muddles the idea that they are at the same meaning level. Second, the elements are scattered and without a center line, which blurs the main concept of gene fusion.

DEGREES OF FREEDOM — ANGULAR



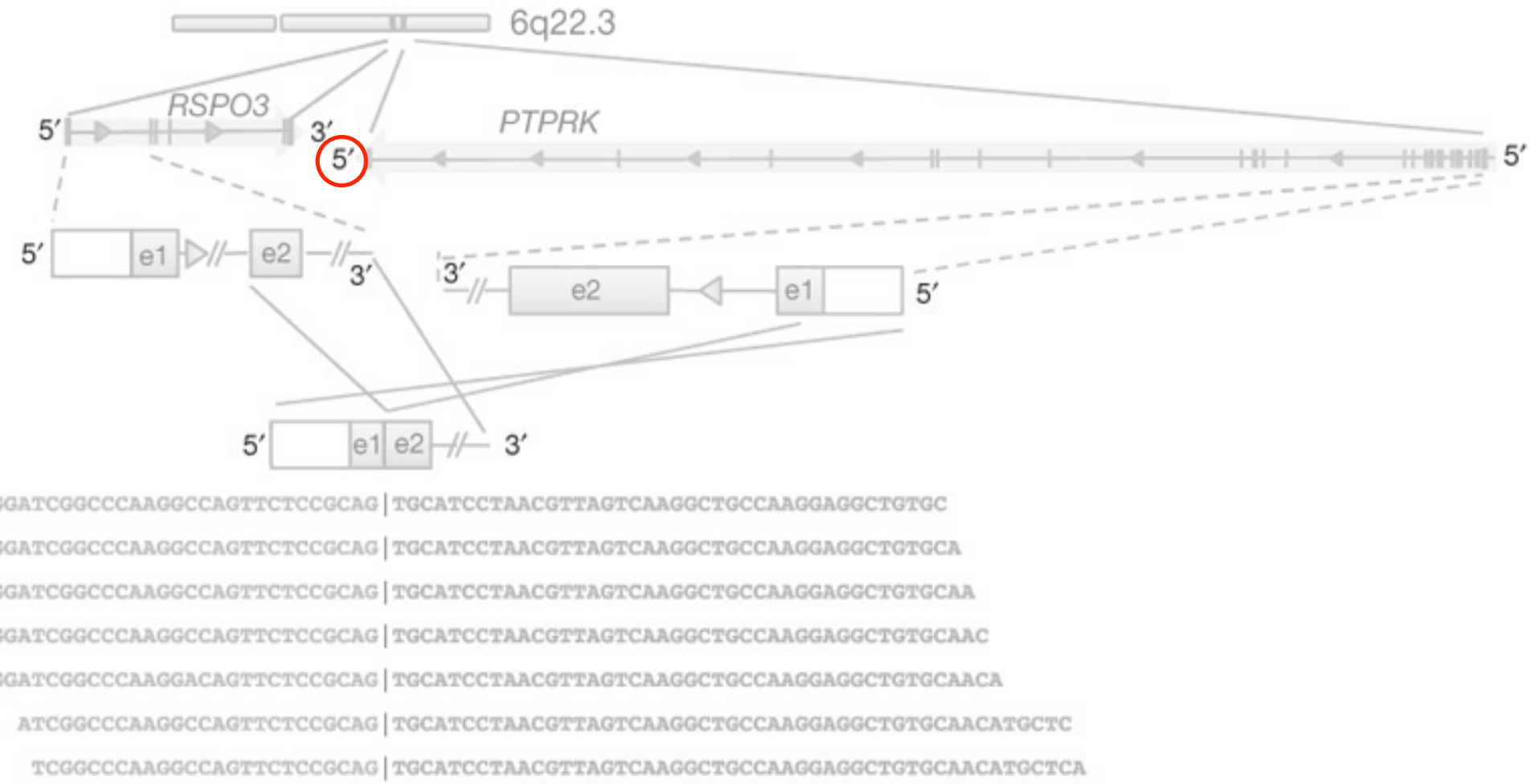
The multitude of angles of the callouts adds unnecessary complexity. When the elements are arranged more consistently, most of the callouts become unnecessary.

REDUNDANCY



The gene model and direction of transcription are repeatedly emphasized in the figure. The direction should be stated once and the model can be dispensed with entirely.

REDUNDANCY — ERRORS CAN BE EASILY MISSED



In the clutter of elements is hiding an error. You probably did not spot it in the original figure because the 5'-3' labels are superfluous. The transcription direction is enough to work out the orientation of each end.

REDUNDANCY — REDUCES EMPHASIS ON DIFFERENCES

Each sequence represents a read that acts as evidence for the observation of gene fusion. By repeating the sequence it makes it difficult to spot minor differences in it, like the protein alignment example from a previous slide. Consider the fact that most readers will not verify that the sequence across reads is conserved - they will simply assume this. Then, if the authors did observe a SNP in a read and wanted to communicate this fact, additional emphasis would need to be placed on this difference to ensure that the reader did not miss it.

DESIGN DECISIONS

what is the core message?

structure and evidence of a gene fusion

what is important?

gene name and orientation
location of breakpoint
change in orientation, if any
local sequence context
supporting evidence

what is not important, or peripheral?

gene size
gene location
gene model (learn to let go)

SCALE AND CONTEXT

REFERENCE

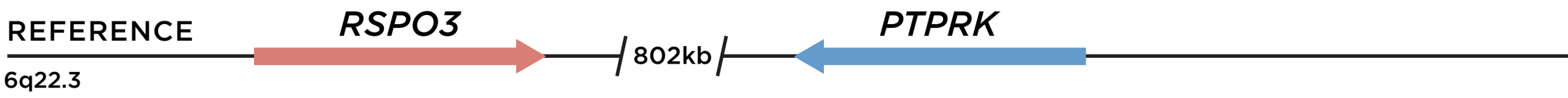
6q22.3

TCTCCTGGGATCGGCCCAAGGCCAGTTCTCCGCAG TGCATCCTAACGTTAGTCAAGGCTGCCAAGGAGGCTGTGCAACATGCTCA

SAMPLE

Emphasizing the idea of a reference and sample makes explaining the concept of gene fusion to a non-specialist audience easier. In many figures in genomics it is not always clear whether the context is the reference or sample sequence. It is not necessary to show the entire chr6 ideogram. In fact, ask yourself whether you're familiar enough with the p/q arm length ratio to distinguish chr5 from chr6. If not, then I suggest that showing the ideogram of chr6 is not useful.

RELEVANT STRUCTURES AND LOCATION

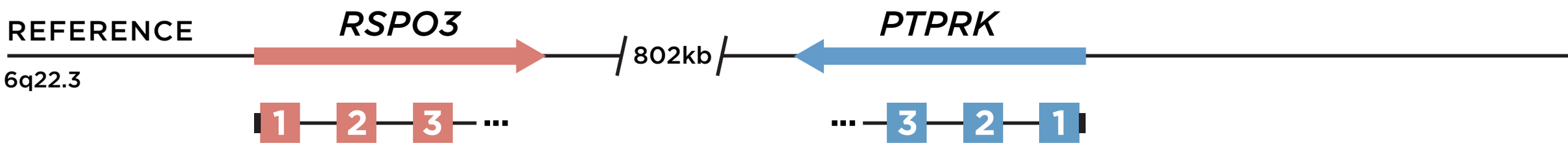


TCTCCTGGGATCGGCCCAAGGCCAGTTCTCCGCAG TGCATCCTAACGTTAGTCAAGGCTGCCAAGGAGGCTGTGCAACATGCTCA

SAMPLE

The exact position of each gene is unimportant, though the distance between them may be.

PROCESS INPUT — ESSENTIAL INTERNALS

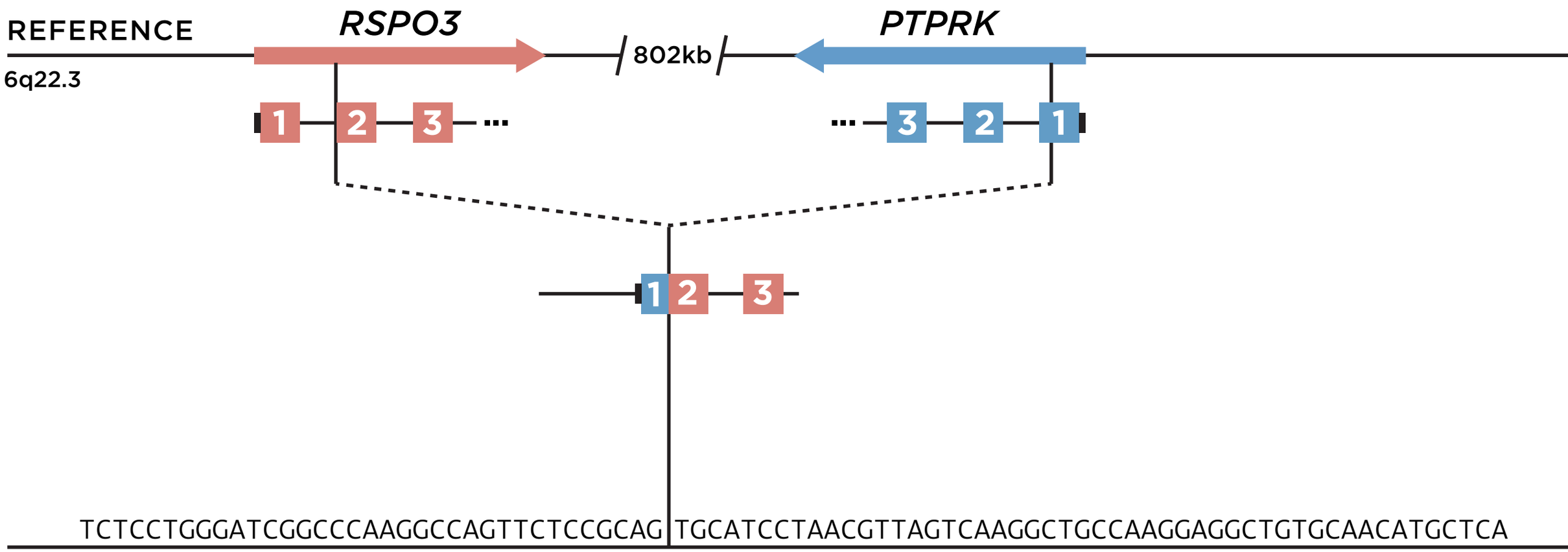


TCTCCTGGGATCGGCCCAAGGCCAGTTCTCCGCAG TGCATCCTAACGTTAGTCAAGGCTGCCAAGGAGGCTGTGCAACATGCTCA

SAMPLE

When exons and their spacing is shown to be uniform, the reader understands that the gene model is a stylized one. The purpose of showing the exons is to support the concept of gene fusion.

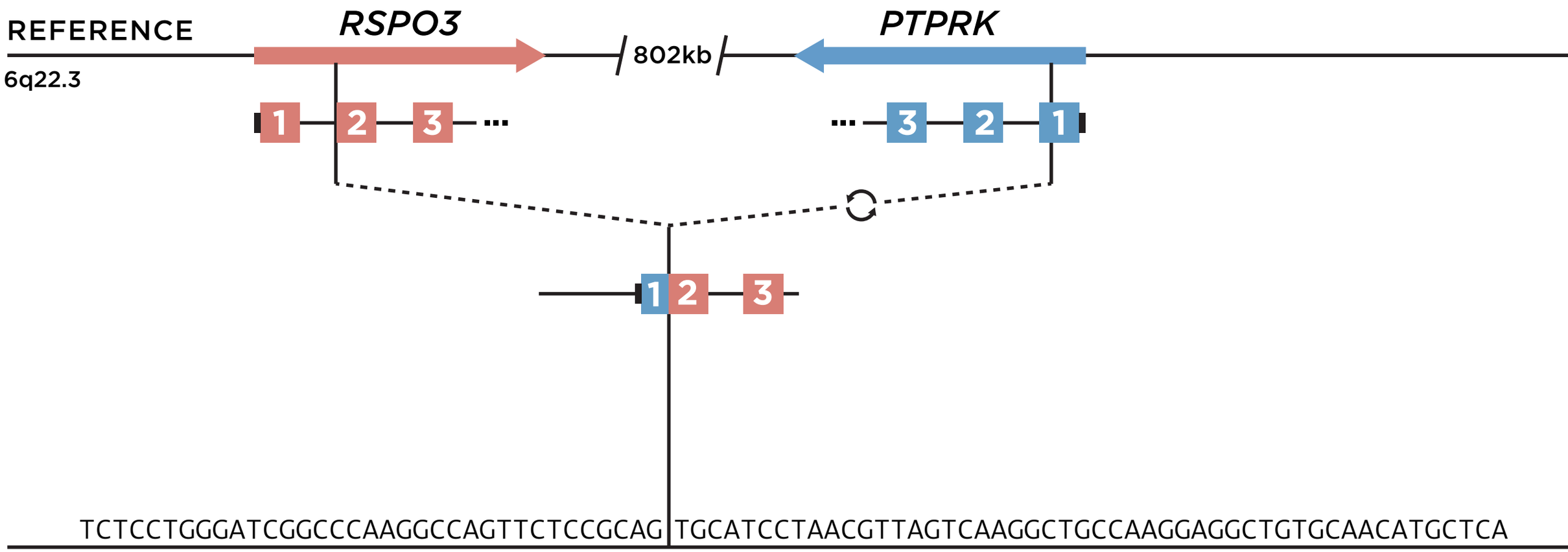
PROCESS OUTPUT — GENE FUSION



This is the critical element of the figure - it connects the reference and sample axes and embodies the concept of gene fusion. The resulting gene fusion product is placed in the center of the figure, giving it emphasis.

visualization: communicating, clearly
MKWEB.BCGSC.CA

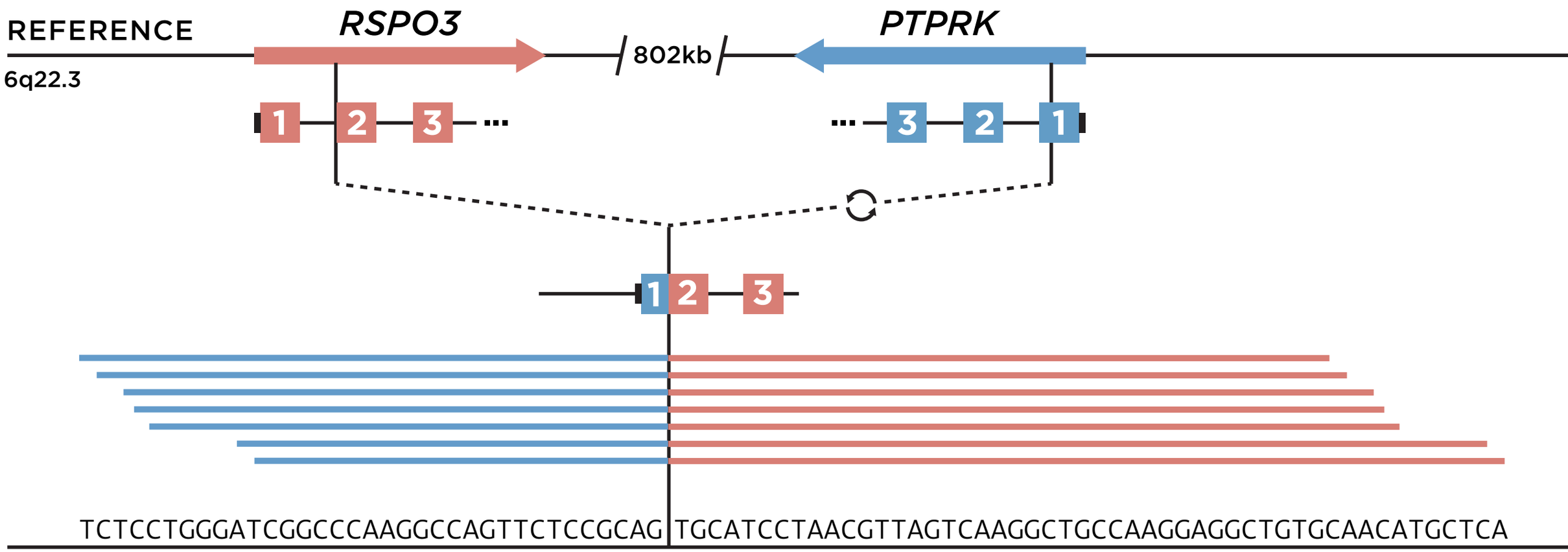
HINTS



A non-specialist reader may not identify that the gene fusion involves an inversion. By using a small hint icon, this fact is quietly emphasized.

visualization: communicating, clearly
MKWEB.BCGSC.CA

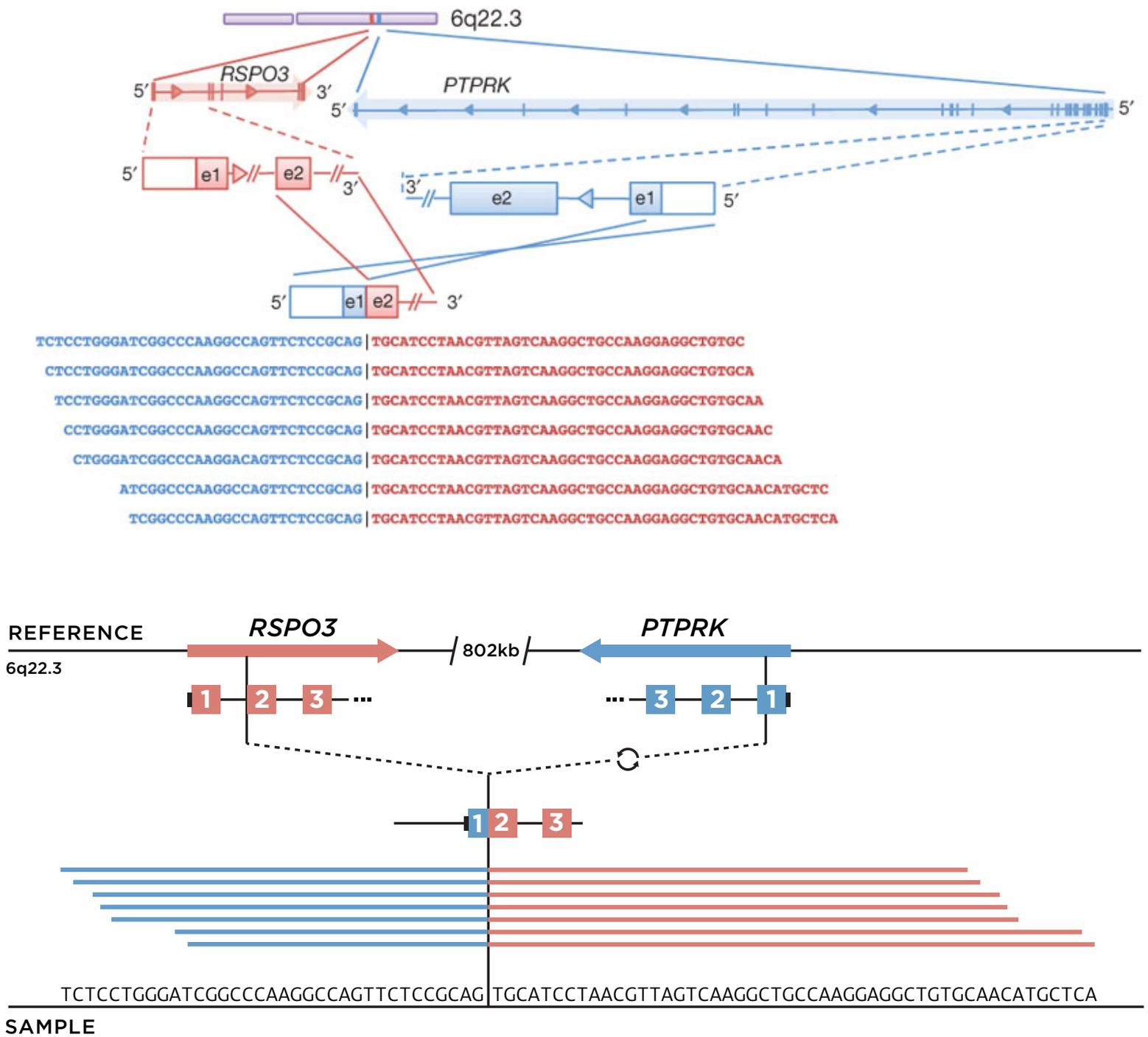
SUPPORT



Rectangular tiles are just as expressive as repeated sequence in communicating the concept of a read. Using tiles allows for additional information, such as SNPs and other sequence variation in the reads.

visualization: communicating, clearly
MKWEB.BCGSC.CA

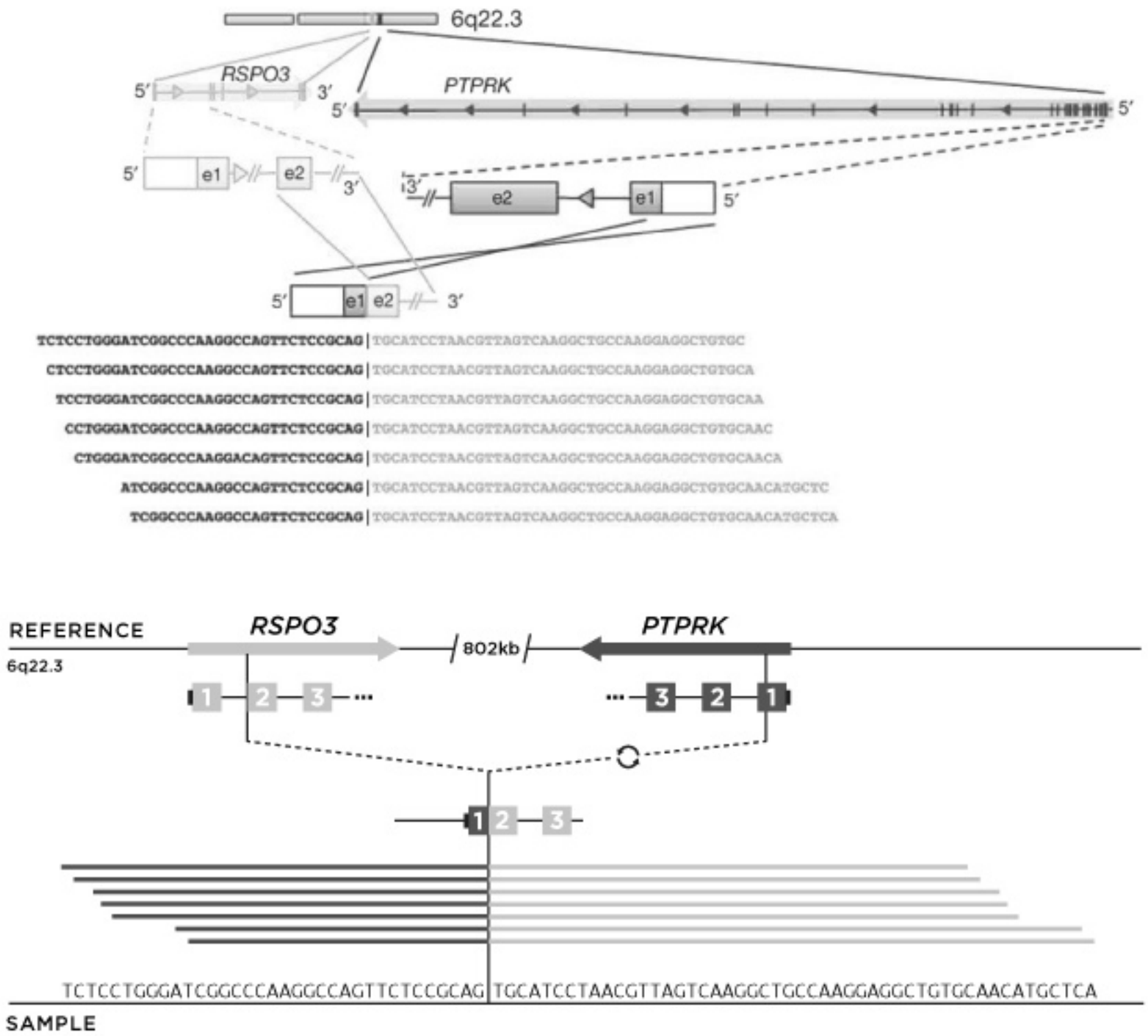
BEFORE & AFTER



Does your figure reproduce well at small size?

Less of an issue in publications, this consideration is important for presentations. The resolution of projects is much lower than printed medium and fine detail can easily be lost to anyone at the back of the room.

GOOD ORGANIZATION OBVIATES NEED FOR COLOR



The original figure uses color to associate elements. This association can be achieved by careful layout and grouping.

The redesign is just as effective in black and white as it is in color.

VISUAL CONCISENESS, CLARITY AND PRECISION

good encoding reveals, good design communicates patterns

visual communication style guide for authors

- hierarchical/themed (callouts⇢arrows⇢heads)
- referenced with usability studies, where available
- linked to good examples in literature
- augmented with templates and tutorials

transitional/strict versions

- do not use curved arrows / use curved arrows sparingly

authors save time

- design decisions are outsourced

readers save time

- figure quality and consistency improved

LITERATURE VIEW & SEARCH, BY FIGURES

journal provides figure rating system

- ●●●●○ informative
- ●●●○○ attractive

figure annotated with ontology (yesterday's breakout)

- form (bar plot)
- content (sequencing)

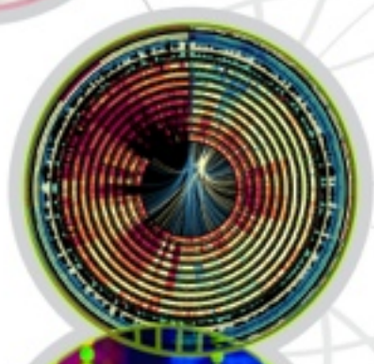
reader can access top quality figures of specific type to learn from good examples

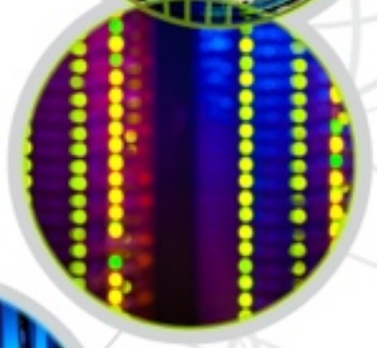
- what are the top 10 scatter plots about sequencing?

GENOMICS

INNOVATION

CANADA'S MICHAEL SMITH
GENOME
SCIENCES
CENTRE





INFORMATICS

SEQUENCING

COMPUTING