Principles of data visualization

Janos Binder

janos.binder@embl.de

EMBL

Agenda

- introduction to visualization
- group work with a figure
- presentations about group work

Acknowledgements

- This slides are based on tutorials of Jessie Kennedy, Martin Krzywinski and Tamara Munzner

YOU WANT TO BE A BETTER COMMUNICATOR

NOT A BETTER ARTIST

Do not rely solely on your personal aesthetic.

Strive for simplicity and clarity.

OBJECTIVE ASPECTS OF DESIGN

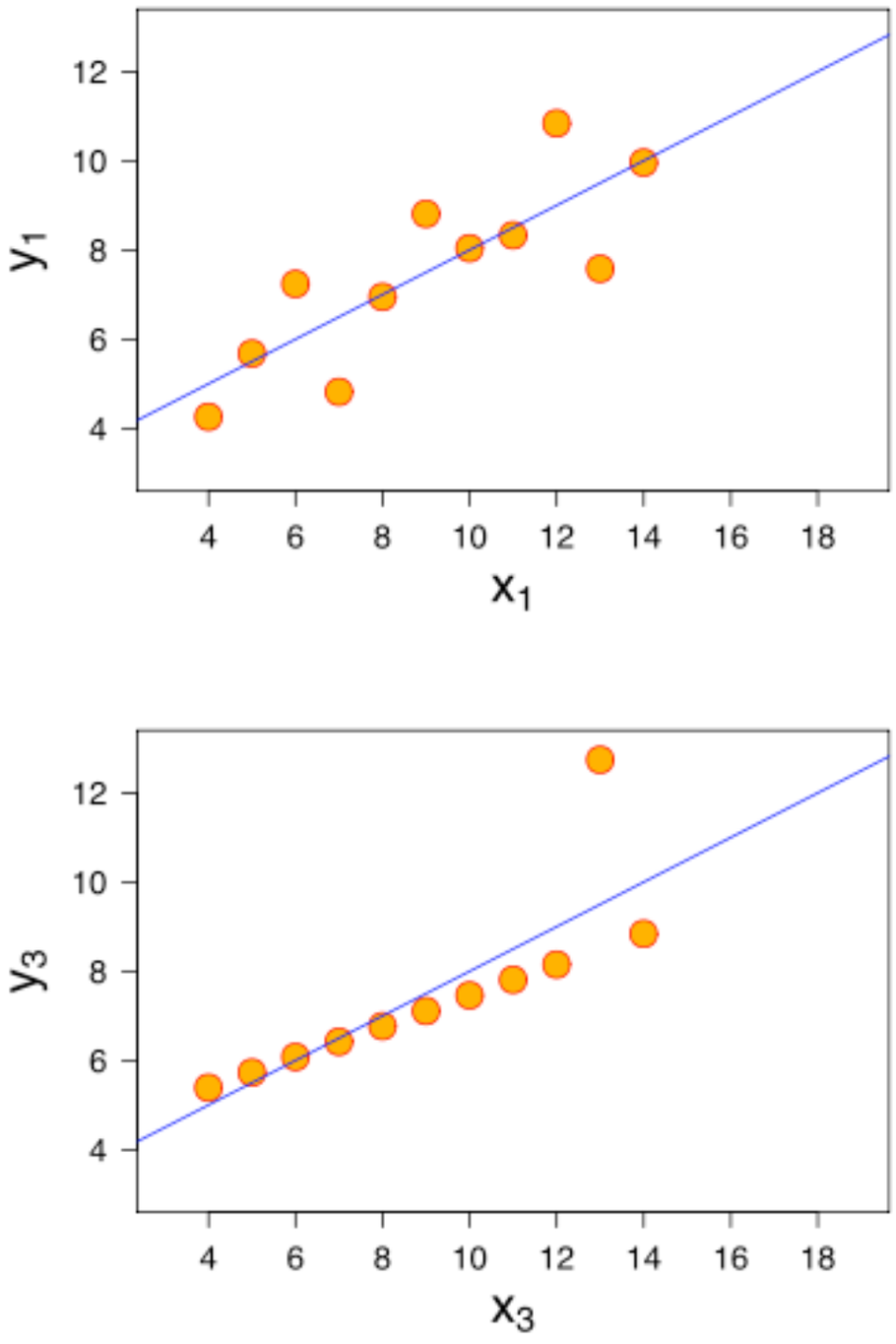
It’s not all about taste.

Why do we visualize?

Anscombe's quartet							
I		II		III		IV	
10	8,04	10	9,14	10	7,46	8	6,58
8	6,95	8	8,14	8	6,77	8	5,76
13	7,58	13	8,74	13	12,74	8	7,71
9	8,81	9	8,77	9	7,11	8	8,84
11	8,33	11	9,26	11	7,81	8	8,47
14	9,96	14	8,1	14	8,84	8	7,04
6	7,24	6	6,13	6	6,08	8	5,25
4	4,26	4	3,1	4	5,39	19	12,5
12	10,84	12	9,13	12	8,15	8	5,56
7	4,82	7	7,26	7	6,42	8	7,91
5	5,68	5	4,74	5	5,73	8	6,89

Same statistical properties: mean(X) = 9, var(X) = 11, mean(Y) = 7.5, var(Y) = 4.12, cor(X,Y) = 0.816, linear regression line Y = 3 + 0.5*X

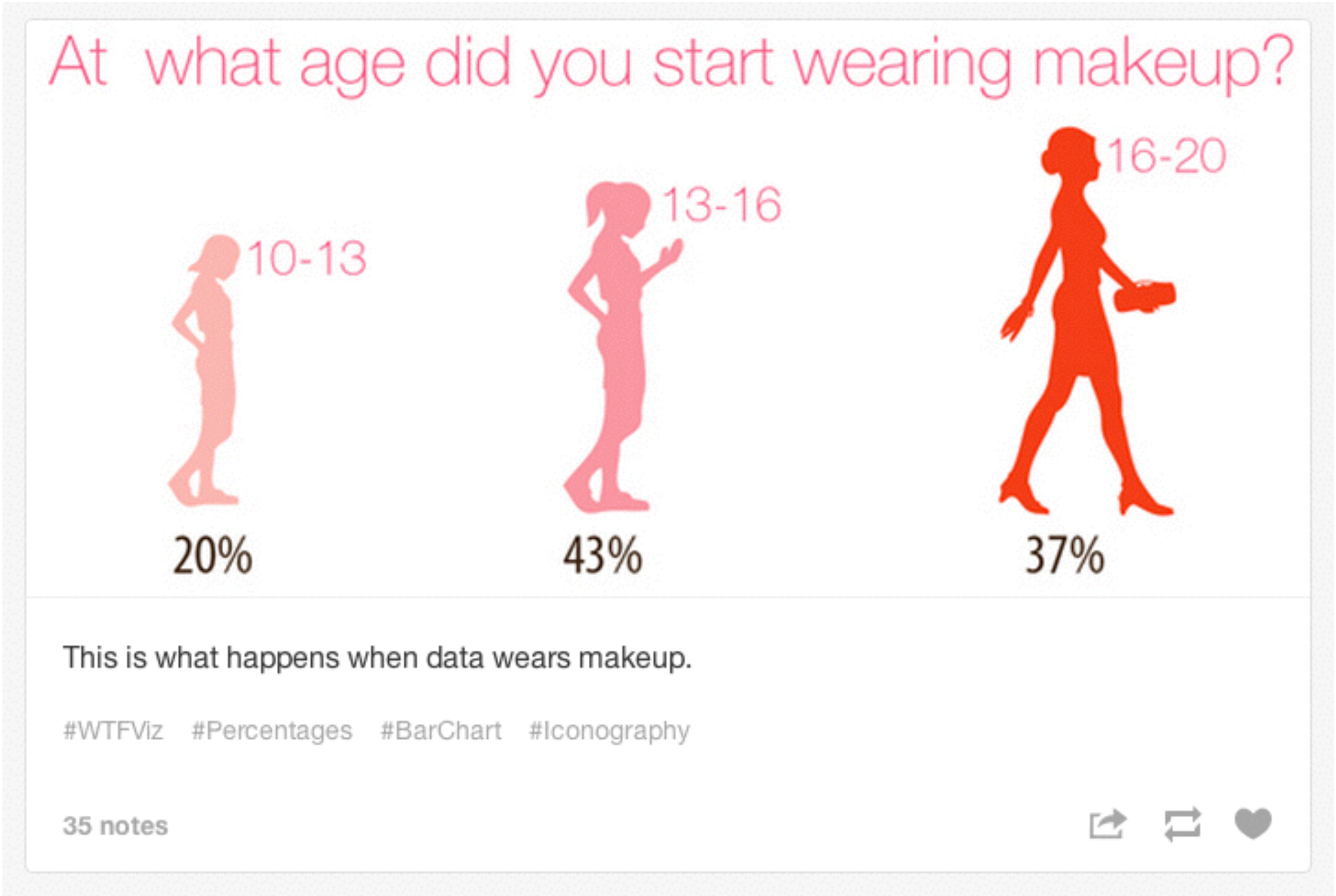
Anscombe's quartet



Is design subjective?

Attractive aspects of proportion, symmetry, color and texture have strong objective foundations.
Few of you would choose the creature on the right as the more attractive.

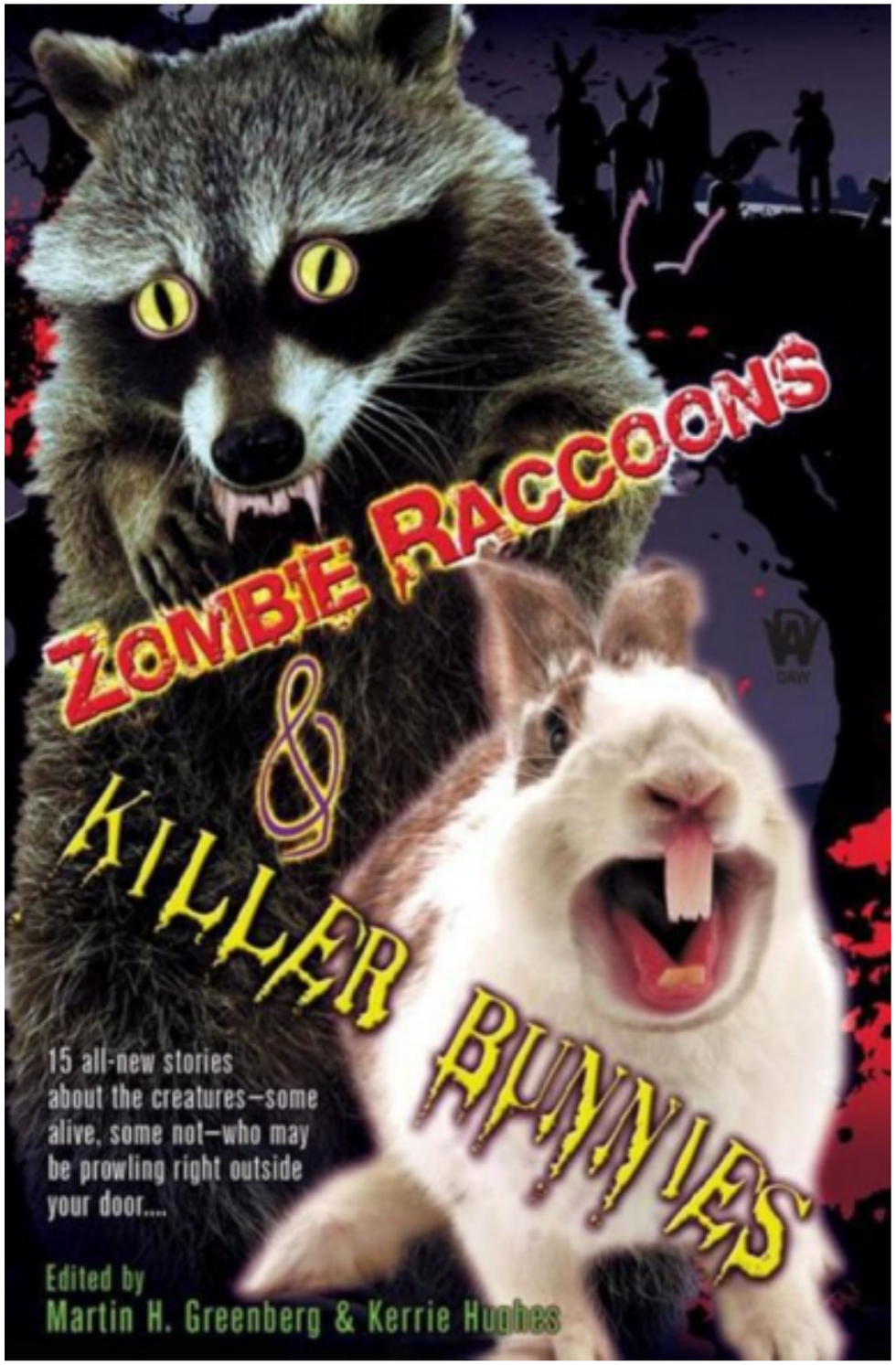
www.wtfviz.net

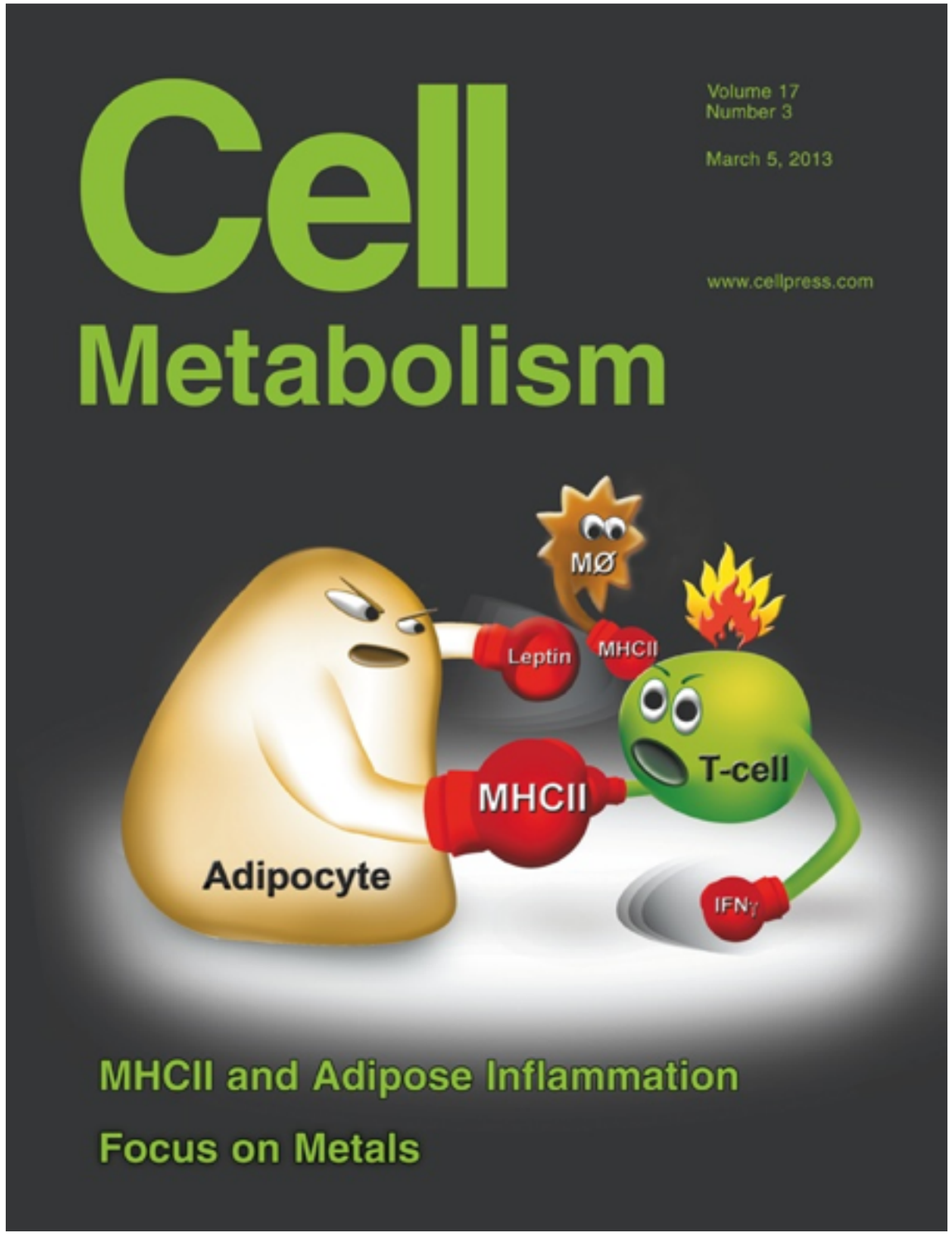


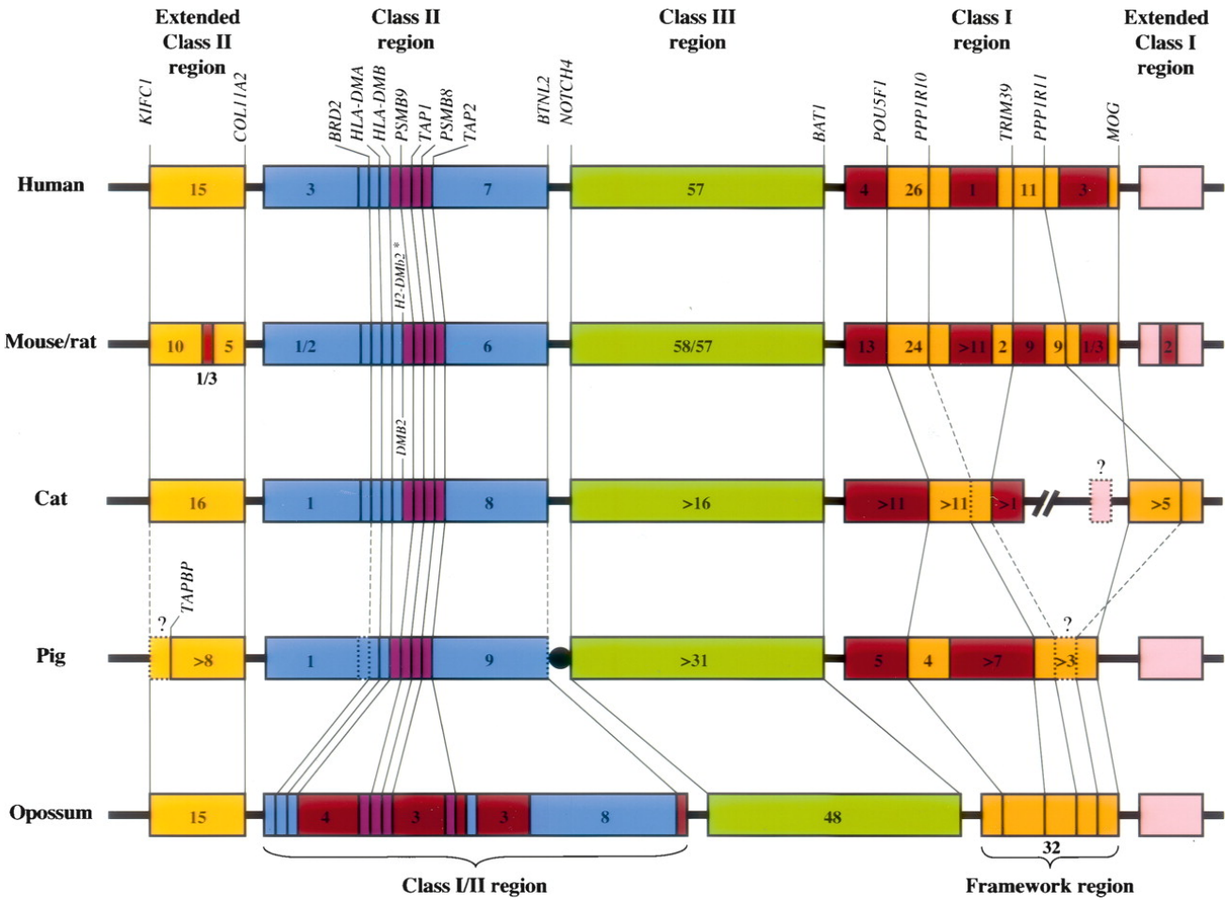
This is what happens when data wears makeup.

#WTFViz #Percentages #BarChart #Iconography

35 notes

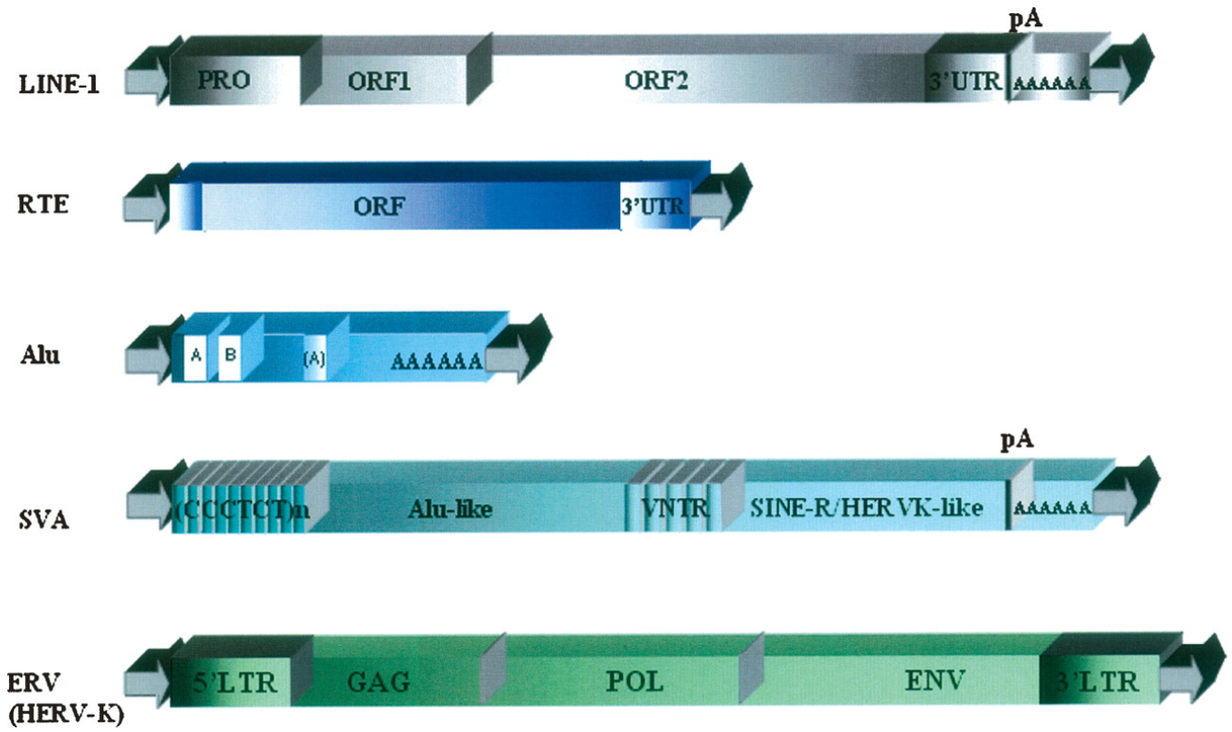

ZOMBIE RACCOONS
&
KILLER BUNNIES
15 all-new stories about the creatures—some alive, some not—who may be prowling right outside your door....
Edited by
Martin H. Greenberg & Kerrie Hughes


Cell
Metabolism
Volume 17
Number 3
March 5, 2013
www.cellpress.com
MØ
Leptin
MHCII
MHCII
T-cell
Adipocyte
IFNγ
MHCII and Adipose Inflammation
Focus on Metals



Excellent organization and consistency. Vertical lines cue continuity. Good use of color.

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



Chartjunk plentiful. Screaming ornamental and redundant elements. Text inconsistent and illegible.

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.

Definition of visualization

Computer-based visualization systems provide visual representations of datasets intended to help people carry out some task more effectively.

Tamara Munzner

Definition of visualization

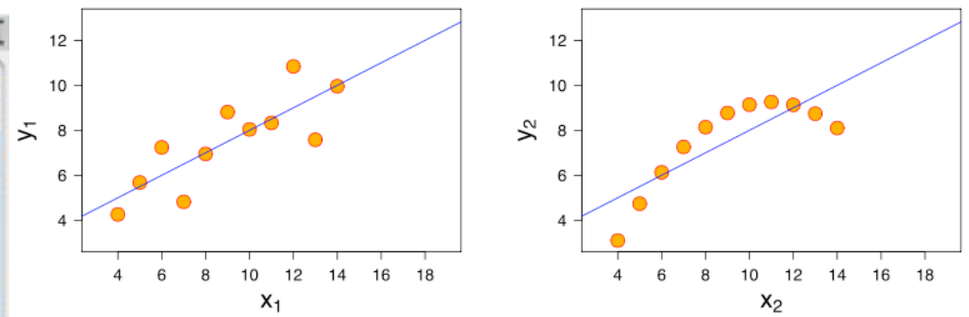
Computer-based visualization systems provide visual representations of datasets intended to help people carry out some task more effectively.

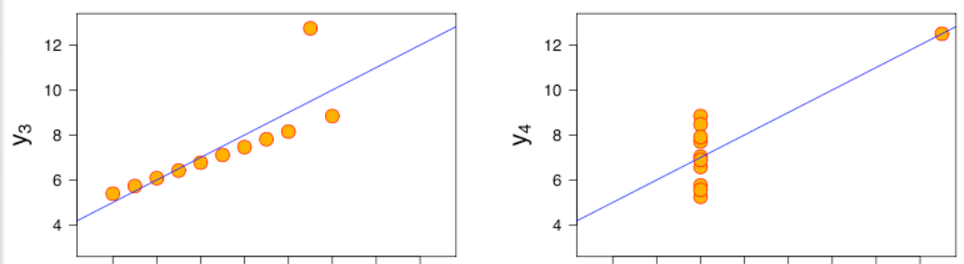
Tamara Munzner

- human in the loop needs the details
- external representation: perception vs cognition
- intended task
- measureable definitions of effectiveness

Data Panel

ID	Function	LPSLL37_1	LPSLL37_1_pvals	LPSLL37_2	LPSLL37_24	LPSLL37_24_pvals
IRAK2	Kinase	2.367	0.251	1.337	-1.553	
NFKB2	Transcription factor	-1.14	0.972	-1.03	1.303	0.807
CXCL2	Chemokine	1.853	0.376	4.111	-1.019	0.745
CHUK	Kinase	-1.376	0.373	2.232	1.194	0.387
IL13	Cytokine	-5.961		2.139	-1.236	0.601
RELA	Transcription factor	-1.077	0.564	-1.169	1.943	0.594
IKBKB	**Kinase**	**1.167**	**0.29**	**1.421**	**-1.907**	**0.286**
CCL4	Chemokine	1.254	0.878	-1.052	1.499	0.761
MAP3K7		1.01	0.956	-1.096	1.222	0.8
ICAM1	Adhesion	1.184	0.669	1.537	1.392	0.671
IRF1	Transcription factor	-1.013	0.519	1.416	1.081	0.995
CXCL3	Chemokine	1.7	0.905	1.092	-1.598	0.521
IL12B	Cytokine	-2.448	0.042	-1.473	-2.109	0.08
CCL11	Chemokine	-1.338	0.349	-1.995	-1.785	0.129
MAP3K7IP1	Adaptor					





SCIENCE IS A PROCESS

DESIGN IS A PROCESS

In either, we don’t always know the end product.

But we must understand how we might get there.

you have data

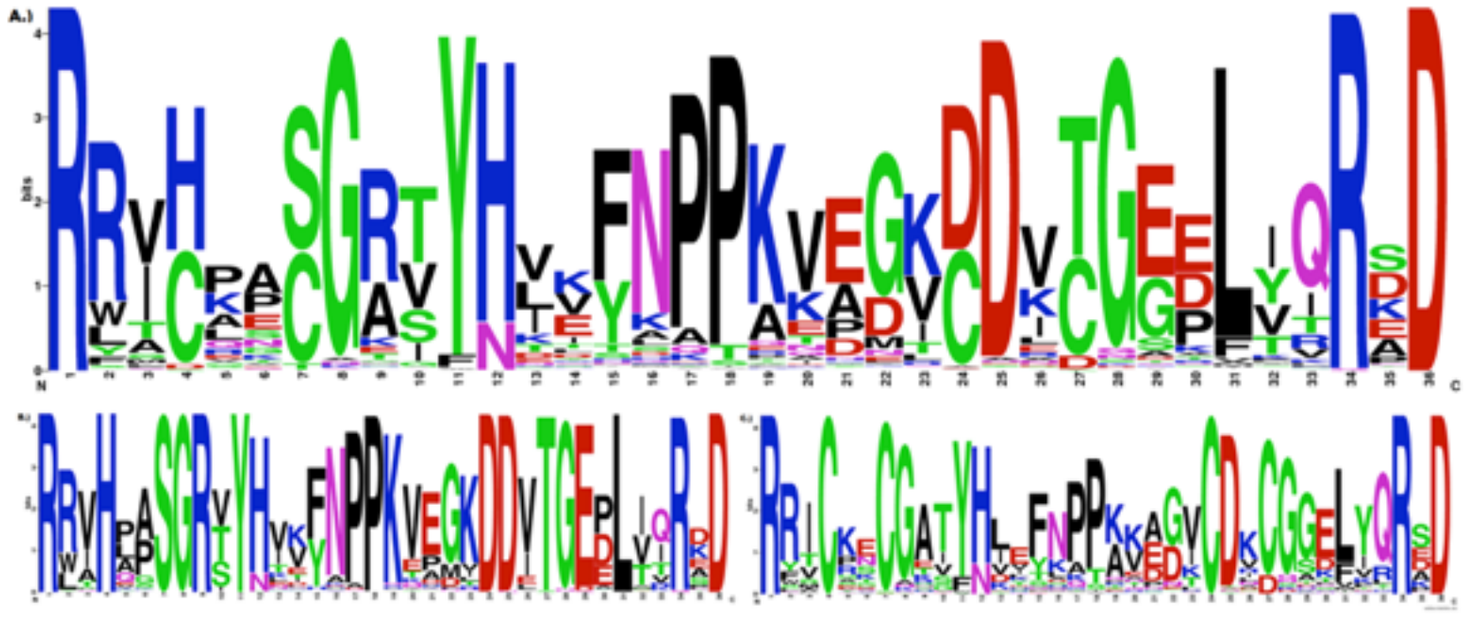
you may have a message

you need to create a figure

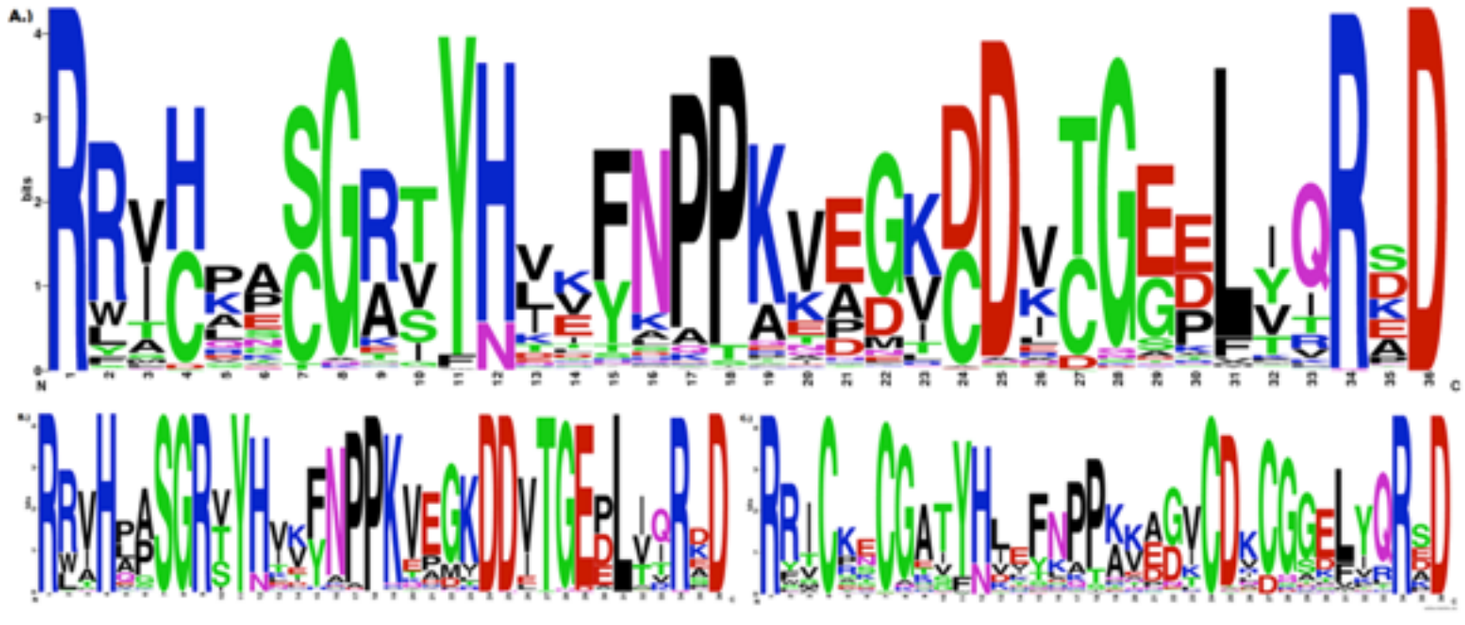
(or think you do)

you select your encoding

e.g. sequence logo



you generate the image with an application that supports the encoding



you write a legend, making up for things that are not obvious

Sequence logos showing the amino acid usage in the adenylate kinase lid (AKL) domain. (A) Across all organisms. (B) from Gram-negative bacteria. (C) from Gram-positive bacteria.

The ADK lid domain structure is universally conserved, but is stabilized in the Gram-negatives by a hydrogen bonding network between residues 4, 7, 9, 24, 27, and 29 (and several other residues in some organisms), while the Gram-positives are stabilized by a bound metal ion, tetrahedrally coordinated by the Cysteines at 4, 7, 24 and 27. **The identities of several other positions (eg 5, 8, 30, 32) are differentially constrained** in each subfamily as well, apparently due to steric requirements of the stabilizing residues.

William Ray, http://www.biovis.net/year/2013/info/redesign-contest

your figure is done

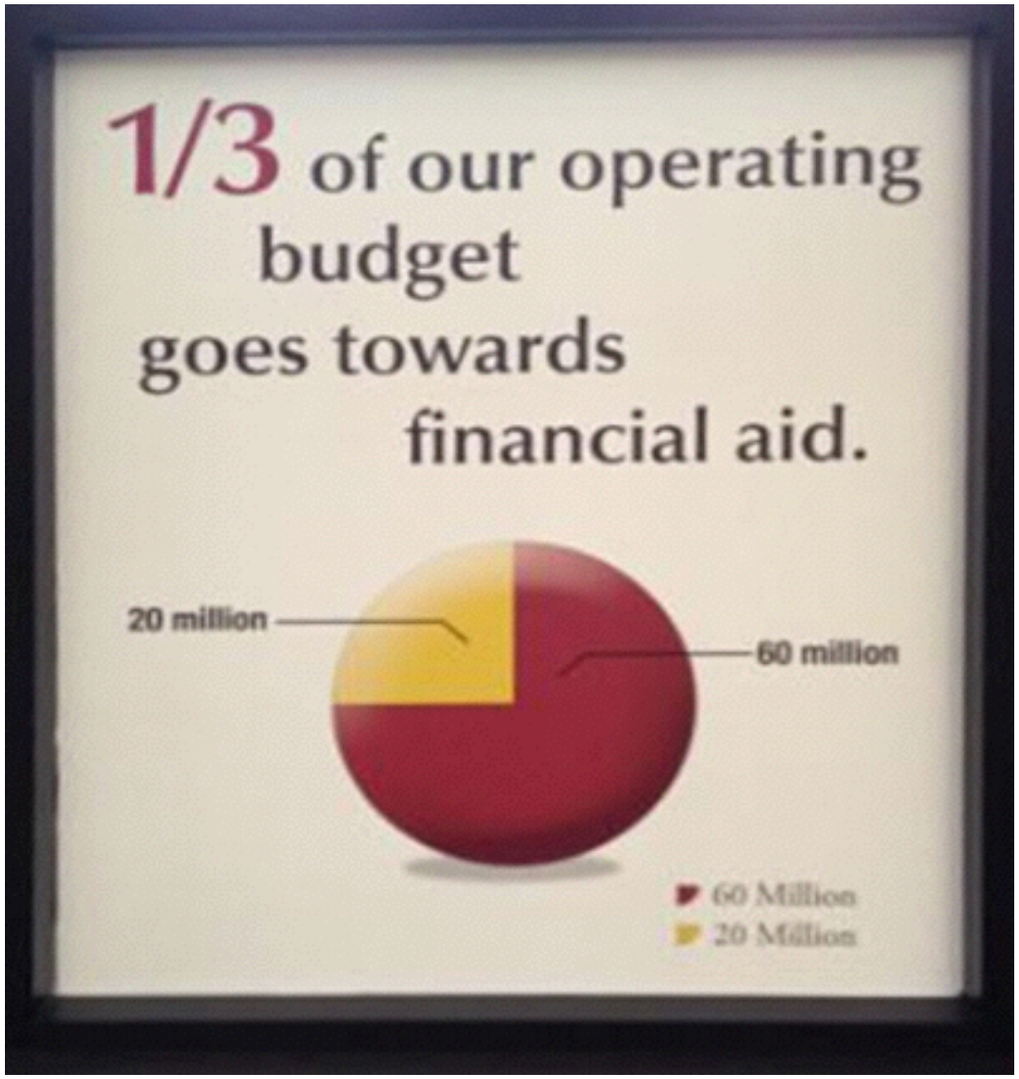
how do you know
whether the creation process
was successful?

CREATE VISUALS WHEN NECESSARY

The desire for a figure is not always proportional to its utility.

lushlush.livejournal.com/190093.html

IS ABSOLUTE ACCURACY ALWAYS IMPORTANT?



via @jayjacobs, https://twitter.com/jayjacobs

NO

accuracy does not always add to utility

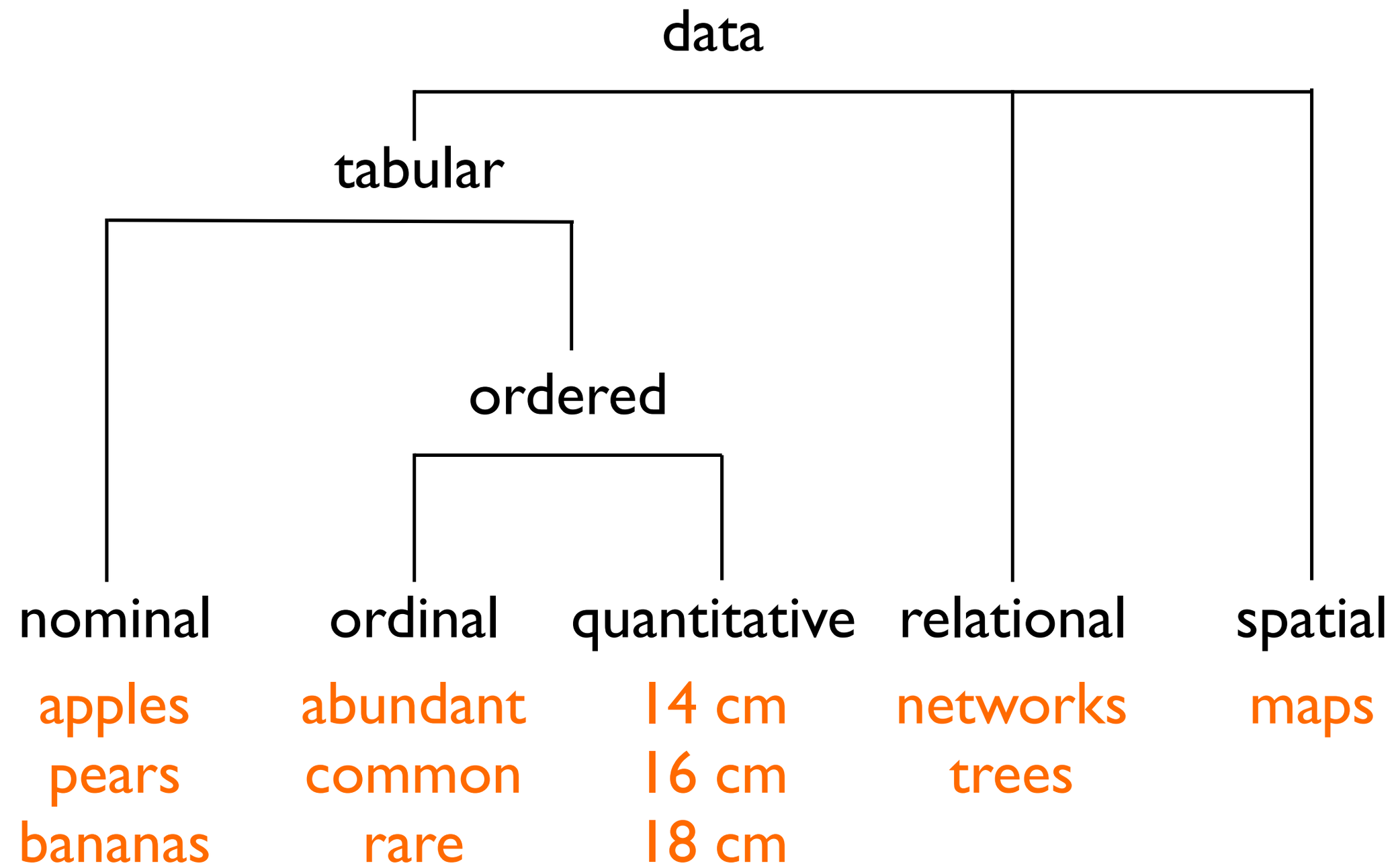
VARIOUS DATA TYPES

What encoding is the best for the data

Data types

apples	abundant	14 cm	networks	maps
pears	common	16 cm	trees	
bananas	rare	18 cm		

Data types


data
tabular
ordered
nominal
ordinal
quantitative
relational
spatial
apples
pears
bananas
abundant
common
rare
14 cm
16 cm
18 cm
networks
trees
maps

Encoding schemes

position

length

density

angle

area

saturation

connection

volume

hue

containment

slope

texture

Accuracy of Quantitative Perceptual Tasks

More accurate

Less accurate

position

length

angle

slope

area

volume

density

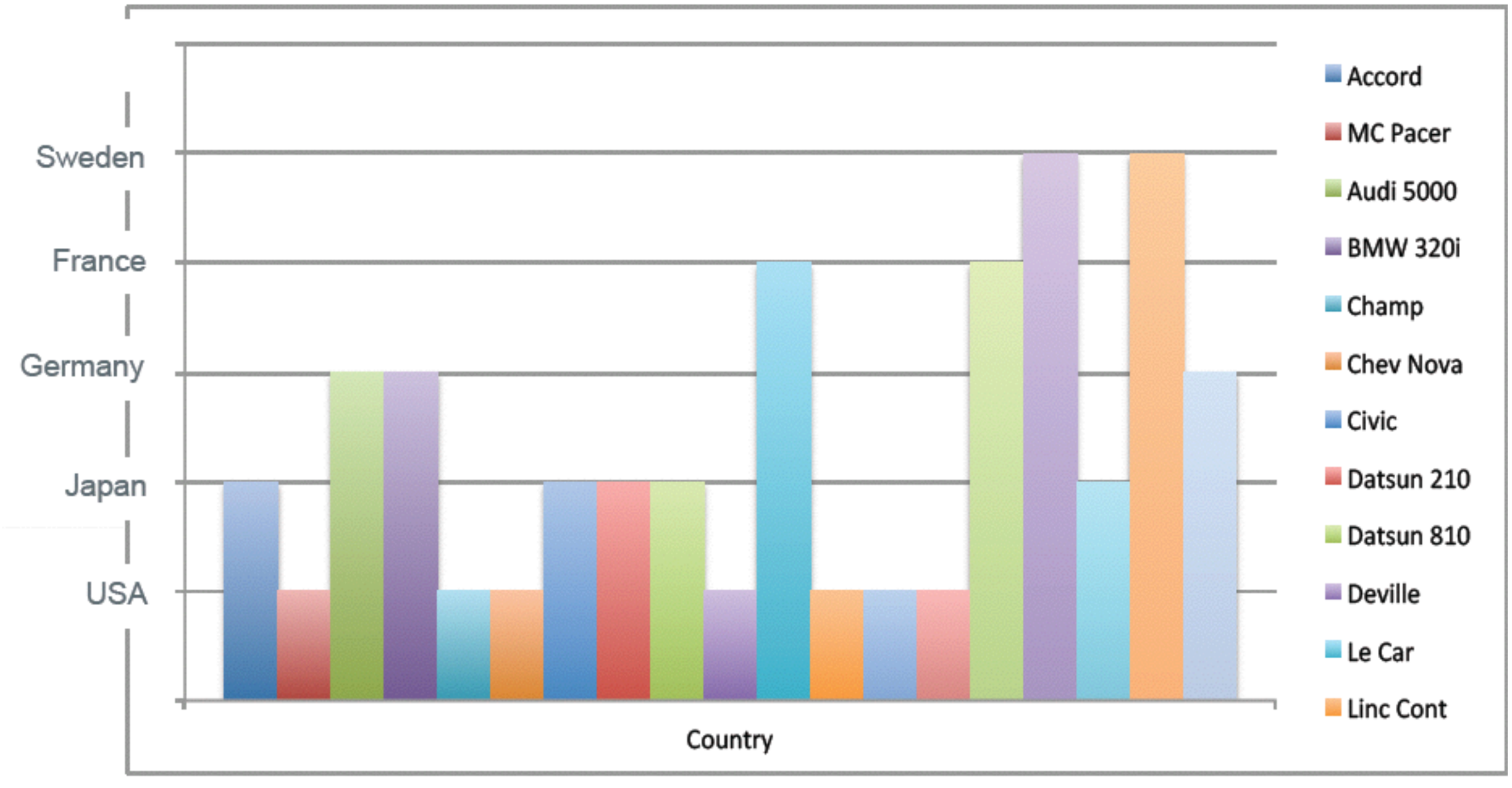
colour

Mapping data types to encoding

Quantitative	Ordinal	Nominal
Position	Position	Position
Length	Density	Hue
Angle	Saturation	Texture
Slope	Hue	Connection
Area	Texture	Containment
Volume	Connection	Density
Density	Containment	Saturation
Saturation	Length	Shape
Hue	Angle	Length
Texture	Slope	Angle
Connection	Area	Slope
Containment	Volume	Area
Shape	Shape	Volume

Expressiveness

- Encodes only the facts



Language of Graphics

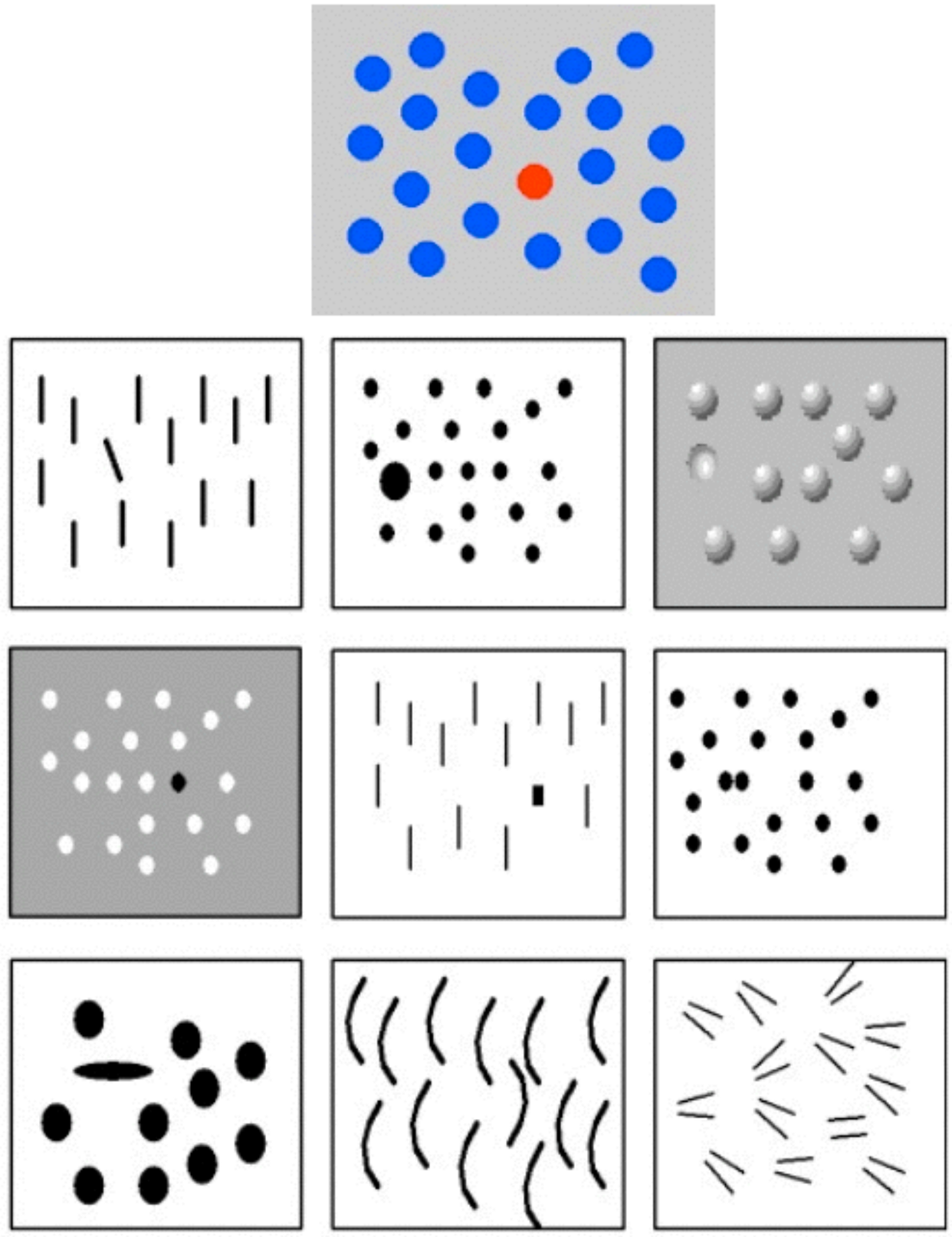
- Graphics can be thought of as forming a sign system:
  - Each mark (point, line, or area) represents a data element.
  - Choose visual variables to encode relationships between data elements
    - difference, similarity, order, proportion
    - only position supports all relationships
- Huge range of alternatives for data with many attributes
  - find images that express and effectively convey the information.

HUMAN PERCEPTION

What features the human brain captures?

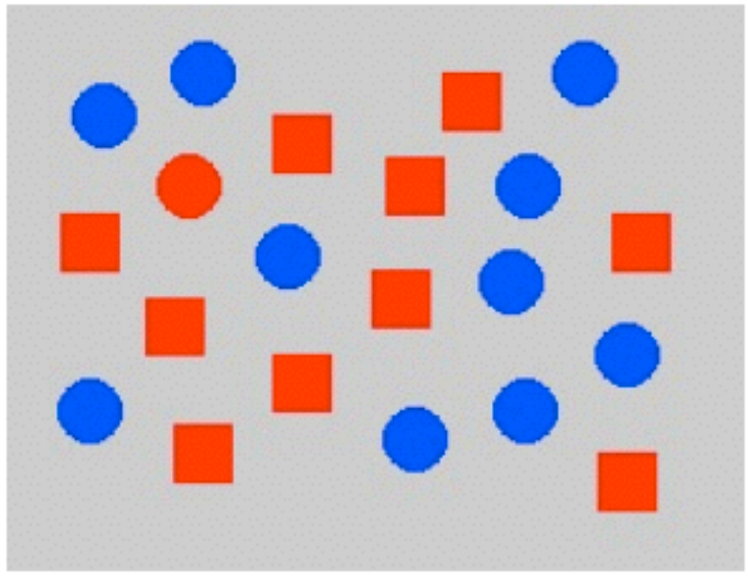
Preattentive Visual Features

- the ability of the low-level human visual system to rapidly identify certain basic visual properties
- a sufficently different item noticed immediately independent of distractor count

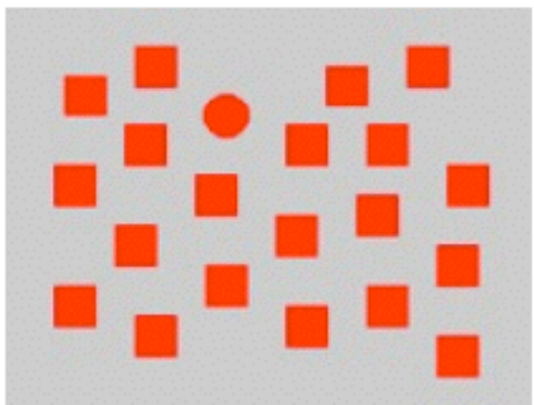


Preattention limits

- only one channel at a time



- within channel, speed depends on which channel and how different item is from the surrounding



Use of preattentive features

- target detection
- boundary detection
- region tracking
- counting and estimation

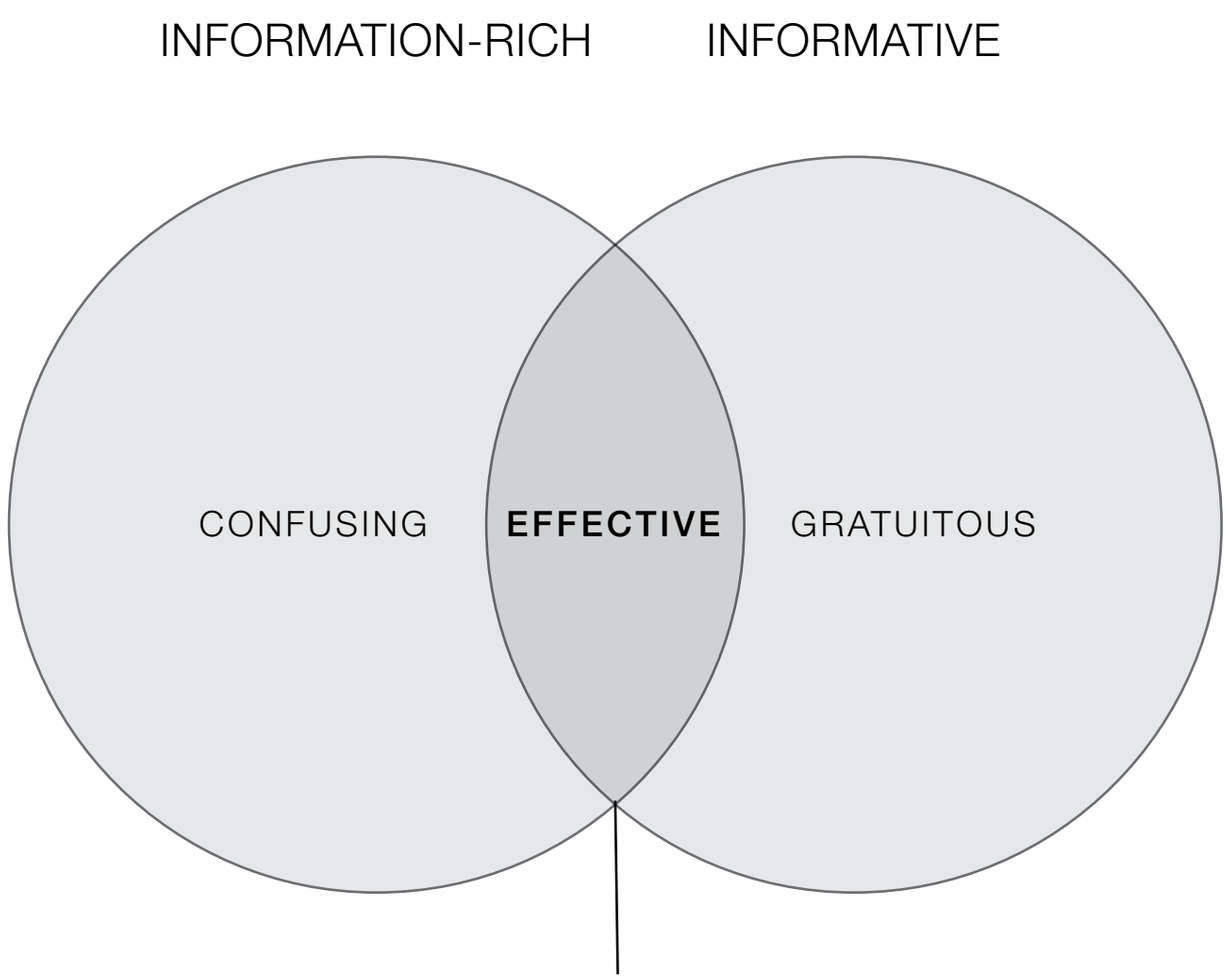
INFORMATIVE + INFORMATION-RICH

In the age of big data, figures should be worth more than 1,000 words.

strive to give your viewer
the greatest number of useful ideas
in the shortest time
with the least ink
in the smallest space

Tufte, E. The Visual Display of Quantitative Information (Graphic Press, Cheshire, Connecticut, USA, 2007).

VISUALIZATION SWEET SPOT

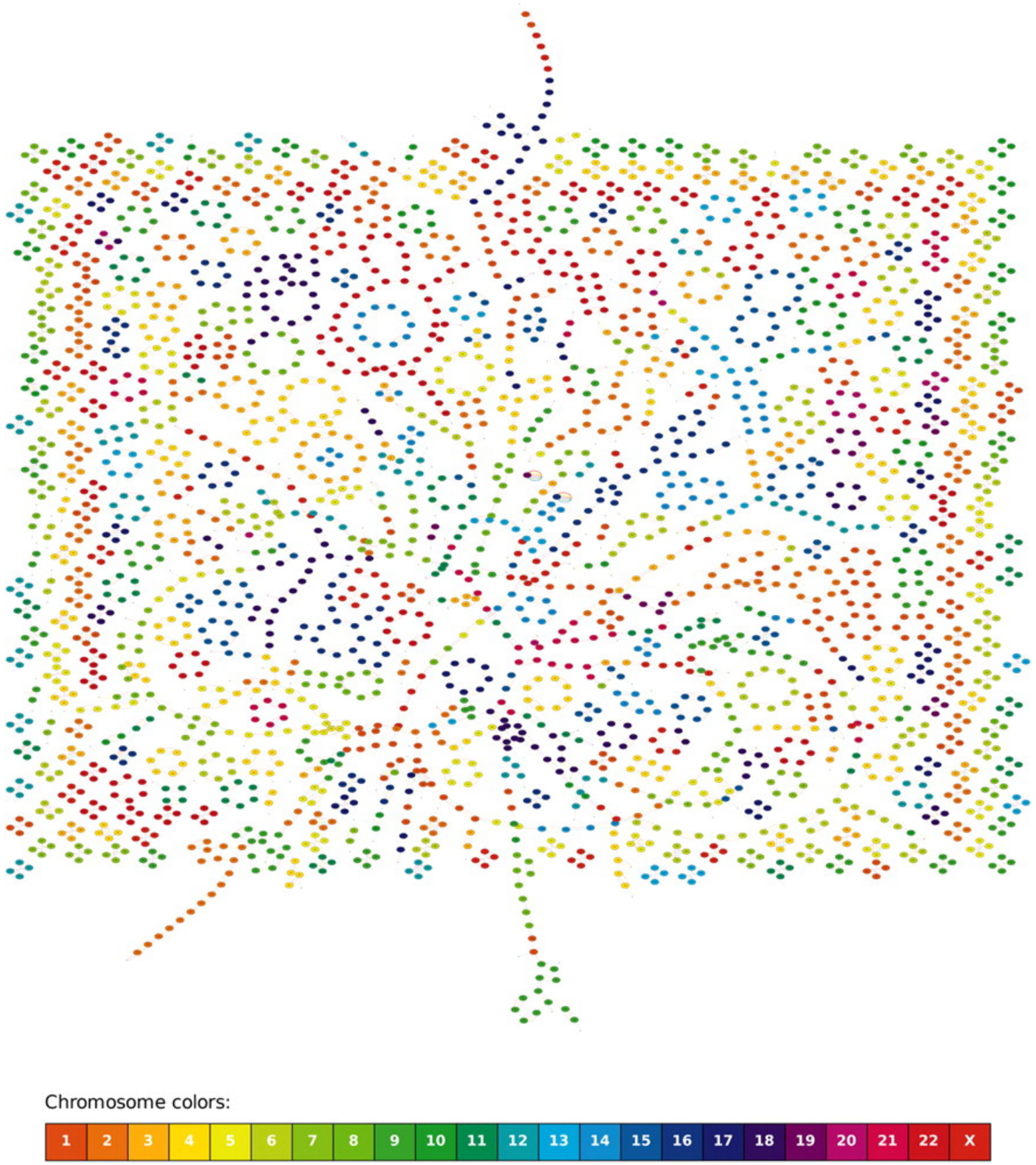


CLEAR MESSAGE

HIGH DATA-TO-INK RATIO

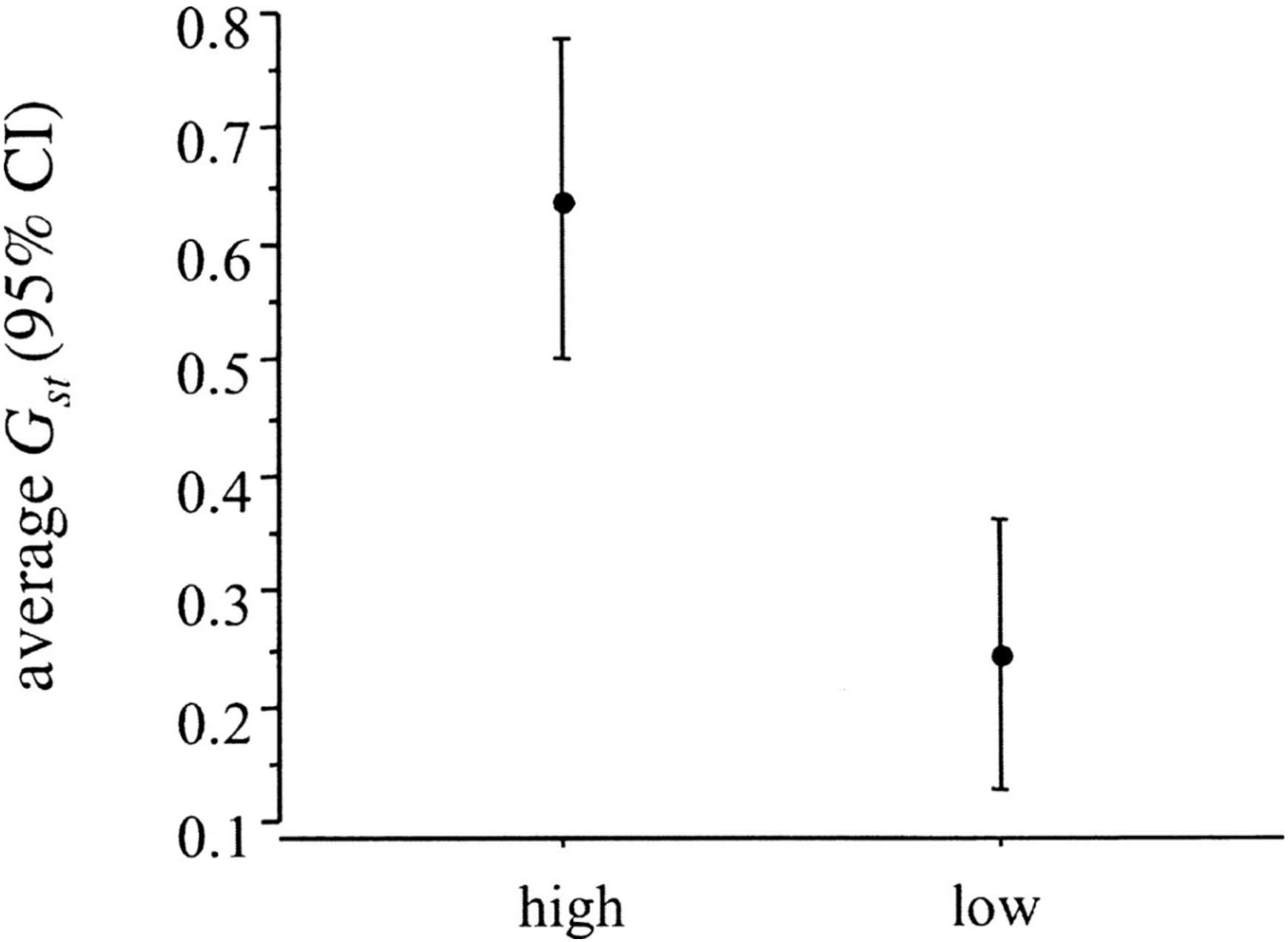
ACCESSIBLE COMPLEXITY

INFORMATION-RICH, NOT INFORMATIVE



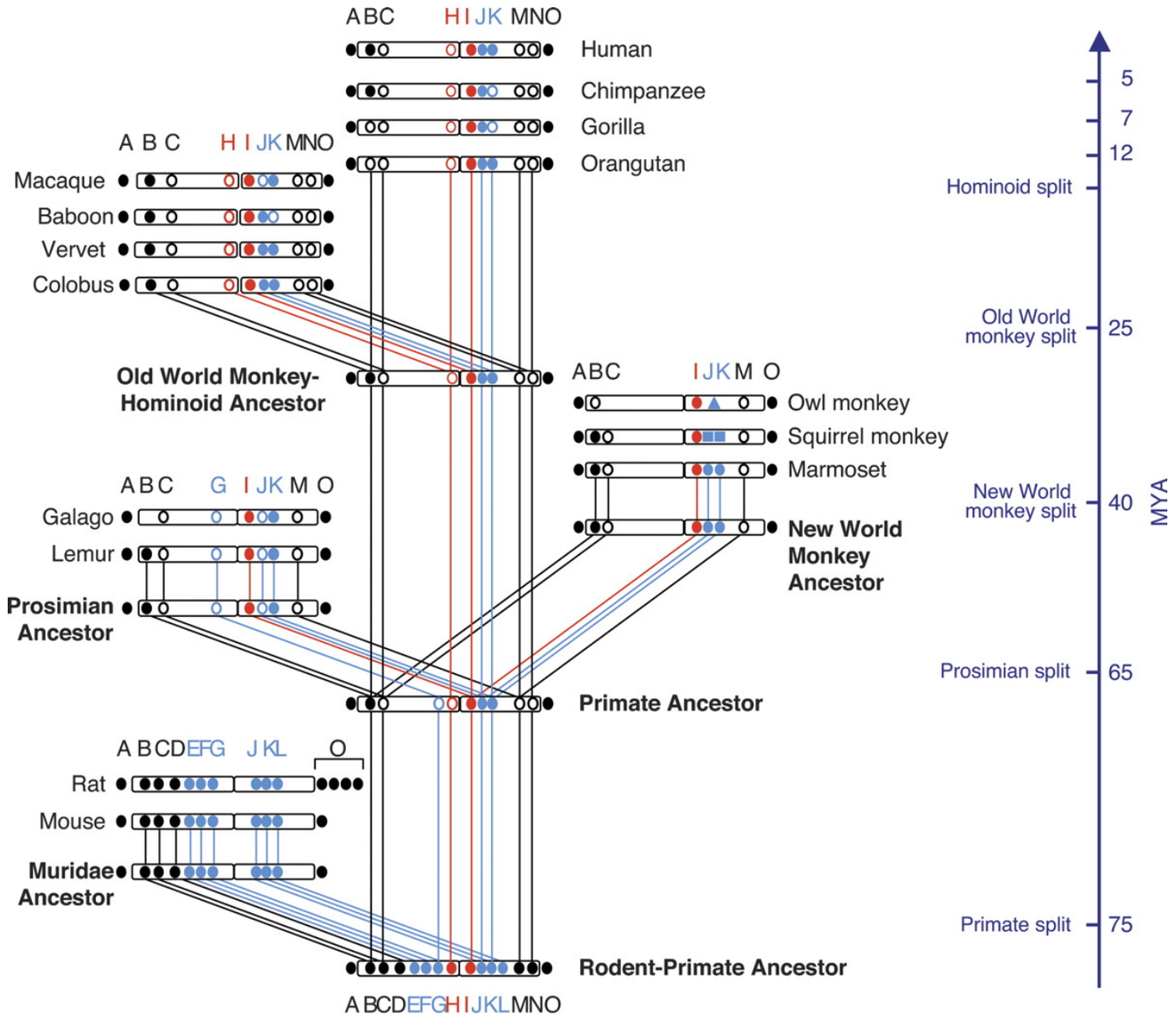
Alekseyev, M.A. and P.A. Pevzner, Breakpoint graphs and ancestral genome reconstructions. Genome Res, 2009. 19(5): p. 943-57.

INFORMATIVE, NOT INFORMATION-RICH



Harr, B., Genomic islands of differentiation between house mouse subspecies. Genome Res, 2006. 16(6): p. 730-7.

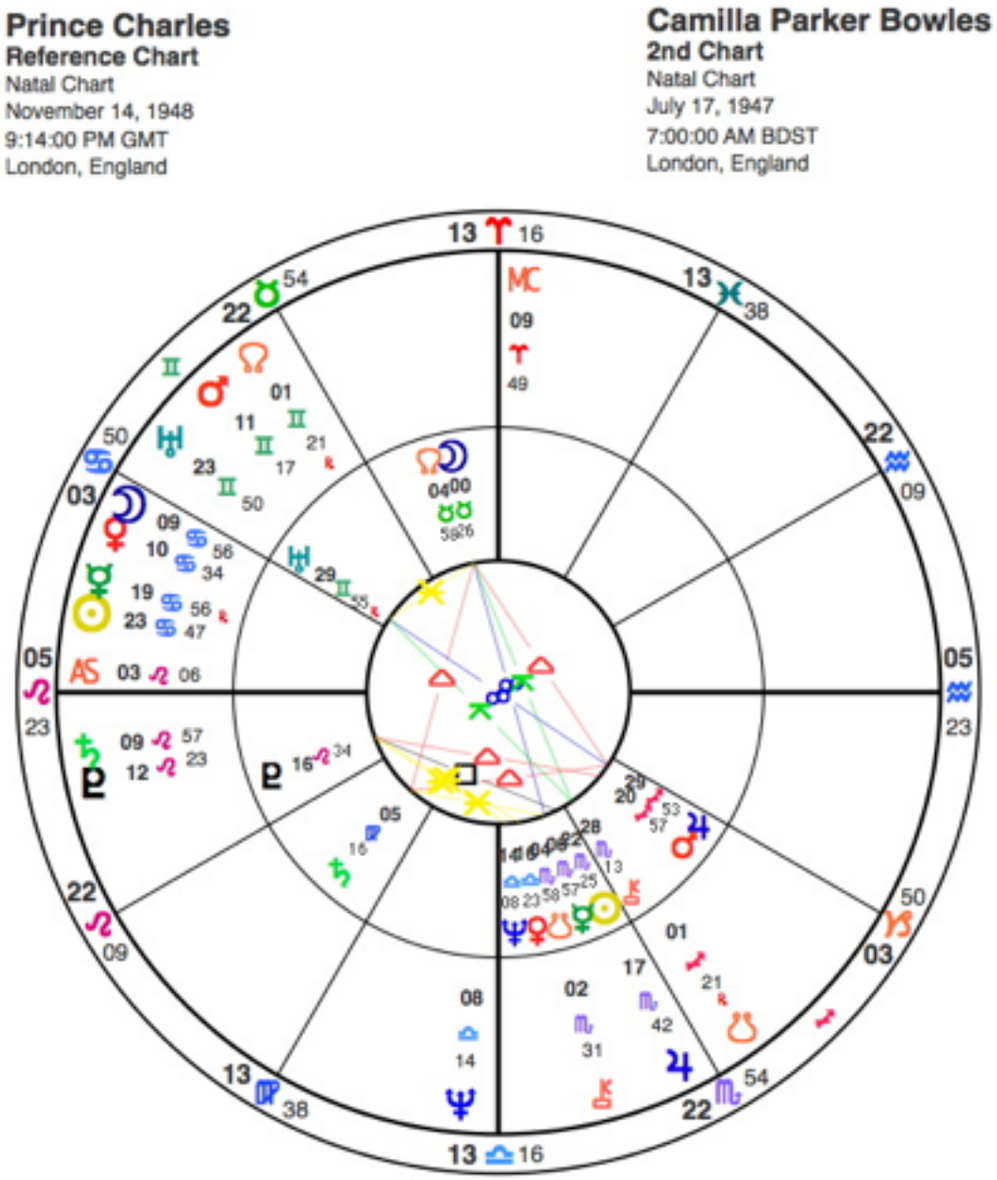
INFORMATION-RICH AND INFORMATIVE



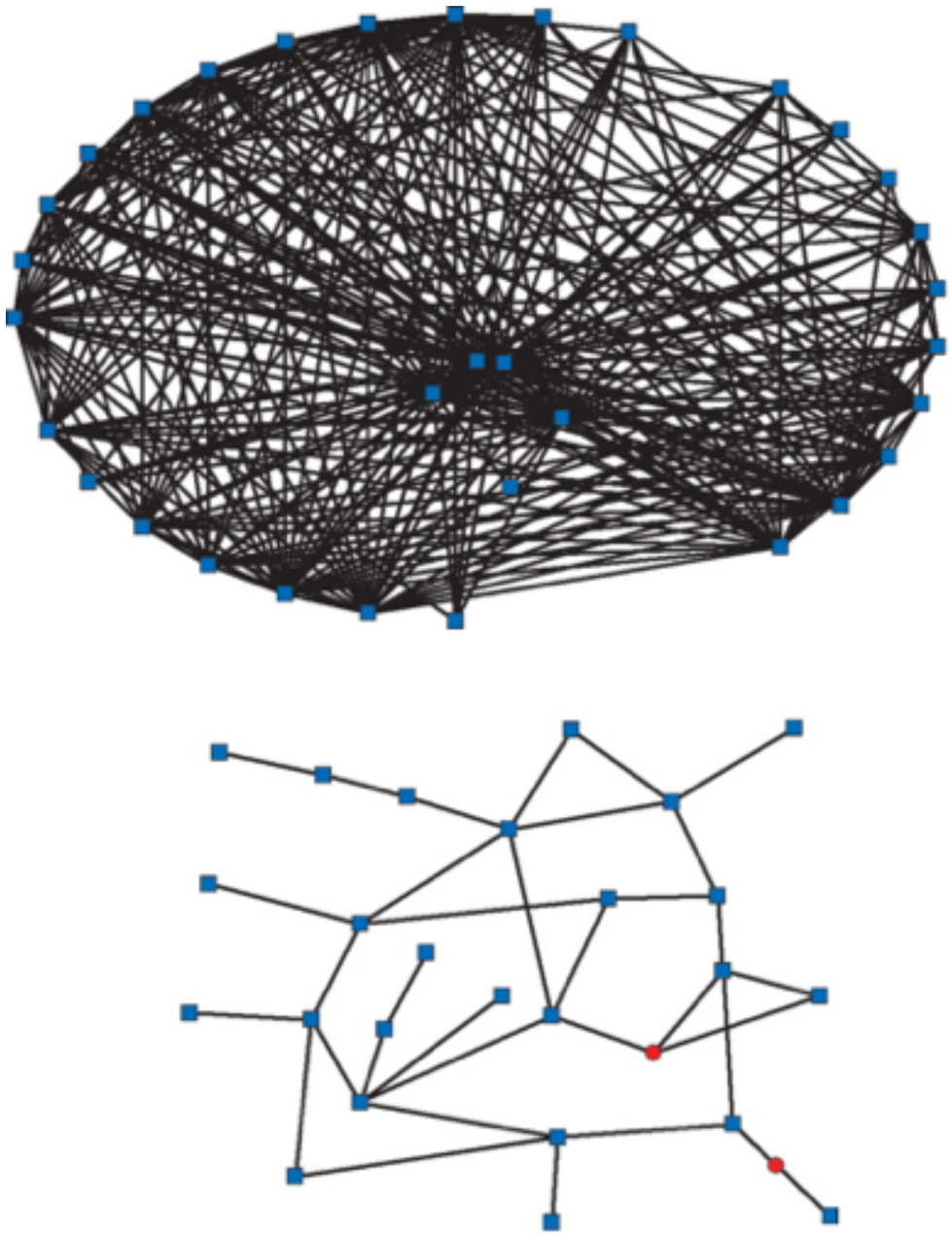
A *WFDC5/Wfdc5*	**D** *WFDC15a/Wfdc15a*	**M** *WFDC15c/Wfdc15c*
B *WFDC12/Wfdc12*	**E, F, G, J, K, L** *SEMG/Svs* family	**N** *WFDC15d/Wfdc15d*
C *WFDC15b/Wfdc15b*	**H, I** Trappin (*PI3*) family	**O** *SLPI/Slpi* family

Hurle, B., W. Swanson, and E.D. Green, Comparative sequence analyses reveal rapid and divergent evolutionary changes of the WFDC locus in the primate lineage. Genome Res, 2007. 17(3): p. 276-86.

JUNK SCIENCE



REAL SCIENCE



(left) Synastry chart. http://sasstrology.com/2011/03/the-astrology-of-marriage-in-the-royal-family-a-suitable-girl-and-the-bit-on-the-side.html
(right) Shakhnovich, B.E. and E.V. Koonin, Origins and impact of constraints in evolution of gene families. Genome Res, 2006. 16(12): p. 1529-36.

don’t merely display data

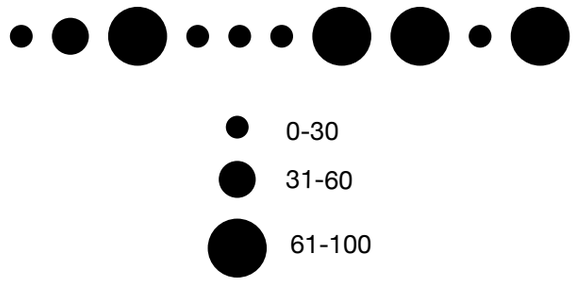
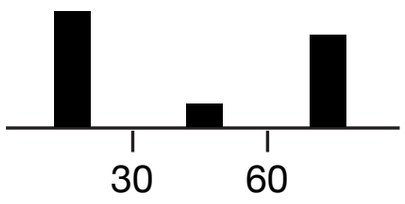
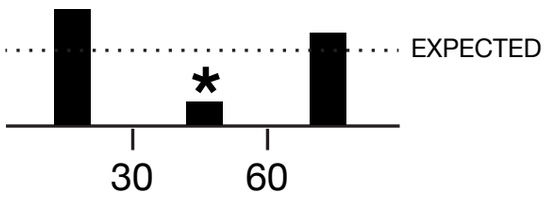
explain it

KNOW YOUR MESSAGE

Stick to it.

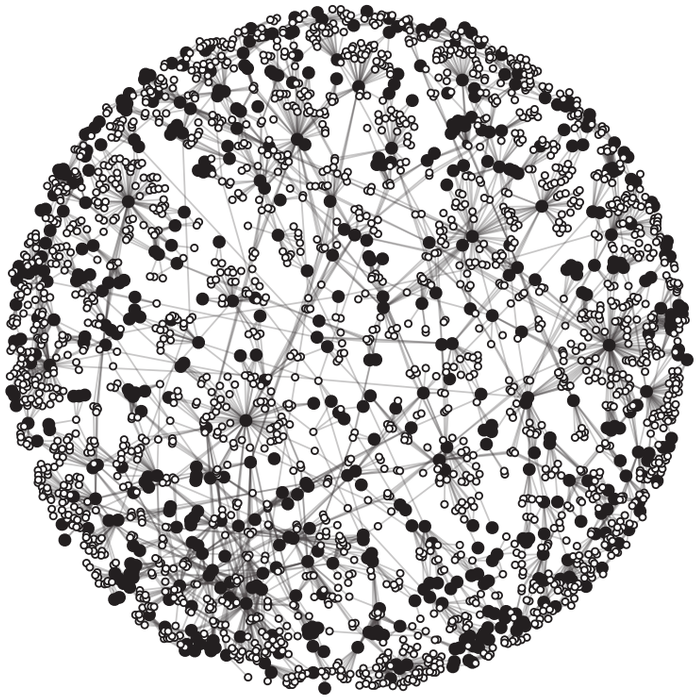

CAUTION
THIS SIGN HAS
SHARP EDGES
DO NOT TOUCH THE EDGES OF THIS SIGN
ALSO, THE BRIDGE IS OUT AHEAD

ACHIEVE FOCUS BY AGGREGATING

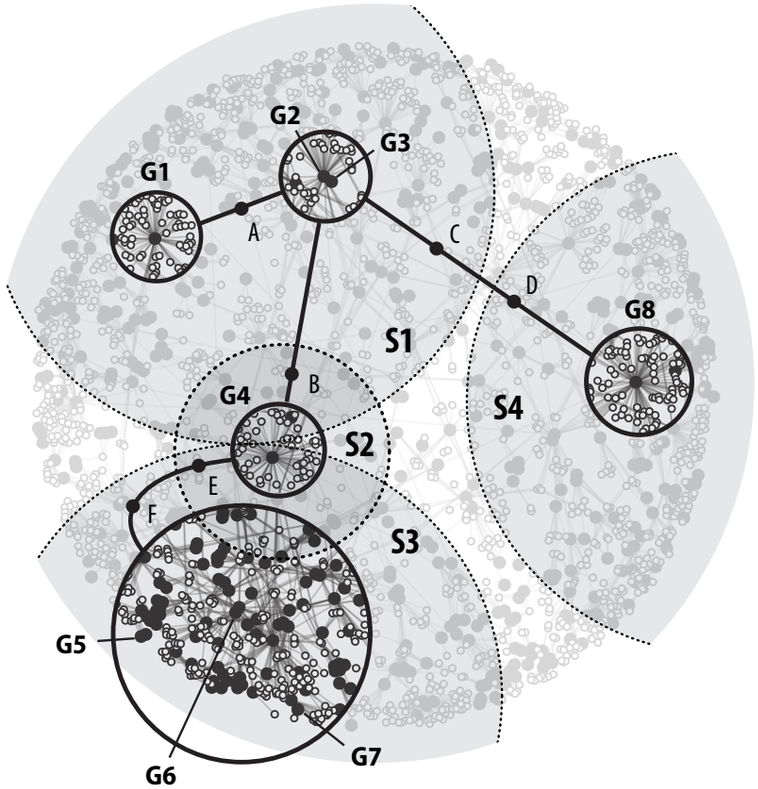
WHAT IS SHOWN?		WHAT IS COMMUNICATED?	WHAT IS INTERPRETED?
RAW DATA	12 54 82 29 25 22 67 61 23 79	NO CLEAR MESSAGE.	UNKNOWN. READER IS ON THEIR OWN.
DISCRETIZED		SCALE	THREE RANGES ARE IMPORTANT. INDIVIDUAL VALUES WITHIN A RANGE ARE NOT.
BINNED		DISTRIBUTION	THERE ARE FEWER MEDIUM-SIZED VALUES.
TREND		SIGNIFICANCE	THERE ARE SIGNIFICANTLY FEWER MEDIUM-SIZED VALUES.

CONTEXT MUST NEVER DILUTE MESSAGE

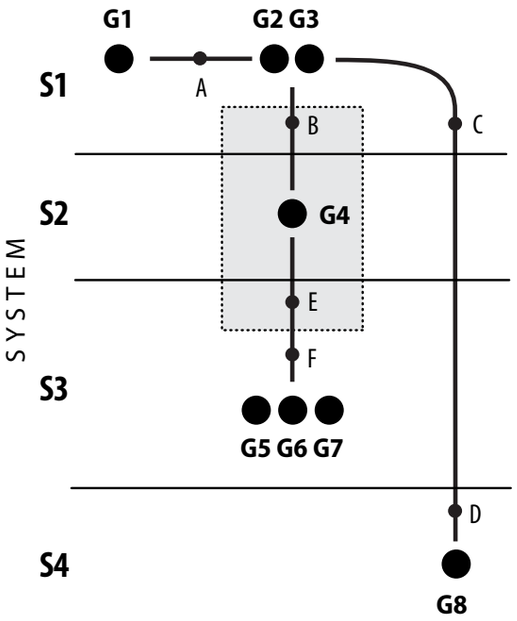
no message



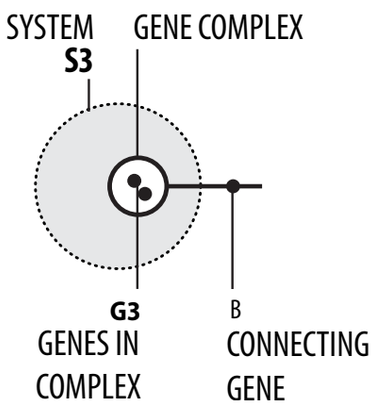
message in context



message in isolation



● GENE ○ DISEASE

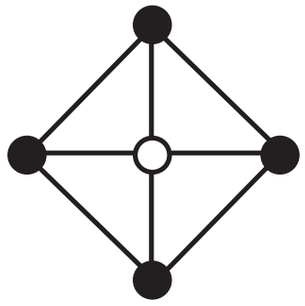


GENE COMPLEX

CONNECTING GENE

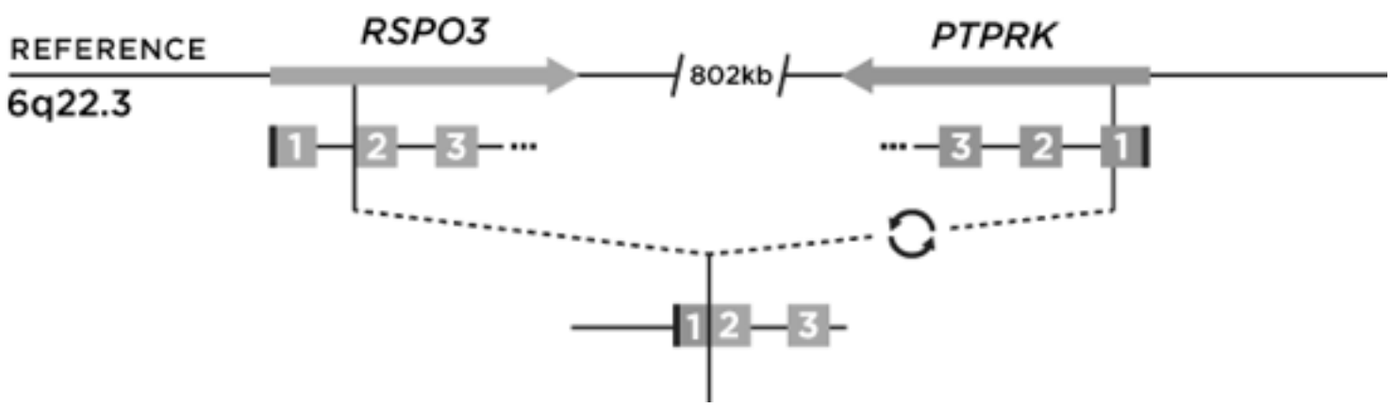
STRIVE FOR CLEAR COMMUNICATION

Revise and redraw.



to explore data, choose effective encoding

to communicate concepts, use effective design



HOW TO APPROACH VISUALIZATION

show the data

induce viewer to think about substance rather than methodology

encourage eye to compare different pieces of data

avoid distorting what the data represents

present many numbers in a small space

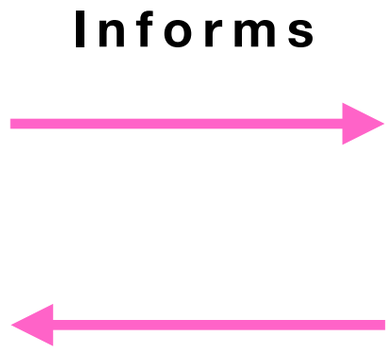
make large data sets coherent

reveal data at several levels of detail – broad overview and fine structure

Tufte, E. The Visual Display of Quantitative Information (Graphic Press, Cheshire, Connecticut, USA, 2007).

TOP-DOWN

redundancy
consistency
conciseness
clarity
focus & emphasis
salience & relevance
truth, accuracy & detail



BOTTOM-UP

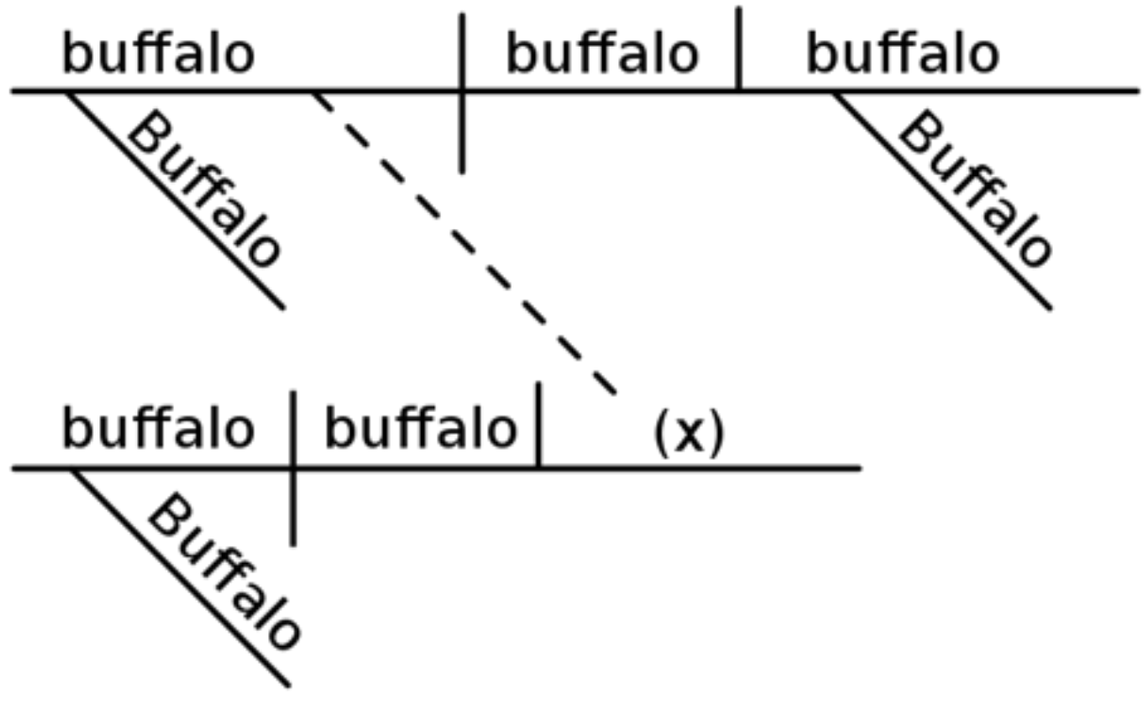
data encoding
symbols
color
typeface
arrows
line weight
alignment

ACCURACY ISN’T EVERYTHING

Correctness does not ensure comprehension.

Buffalo buffalo Buffalo buffalo buffalo buffalo Buffalo buffalo

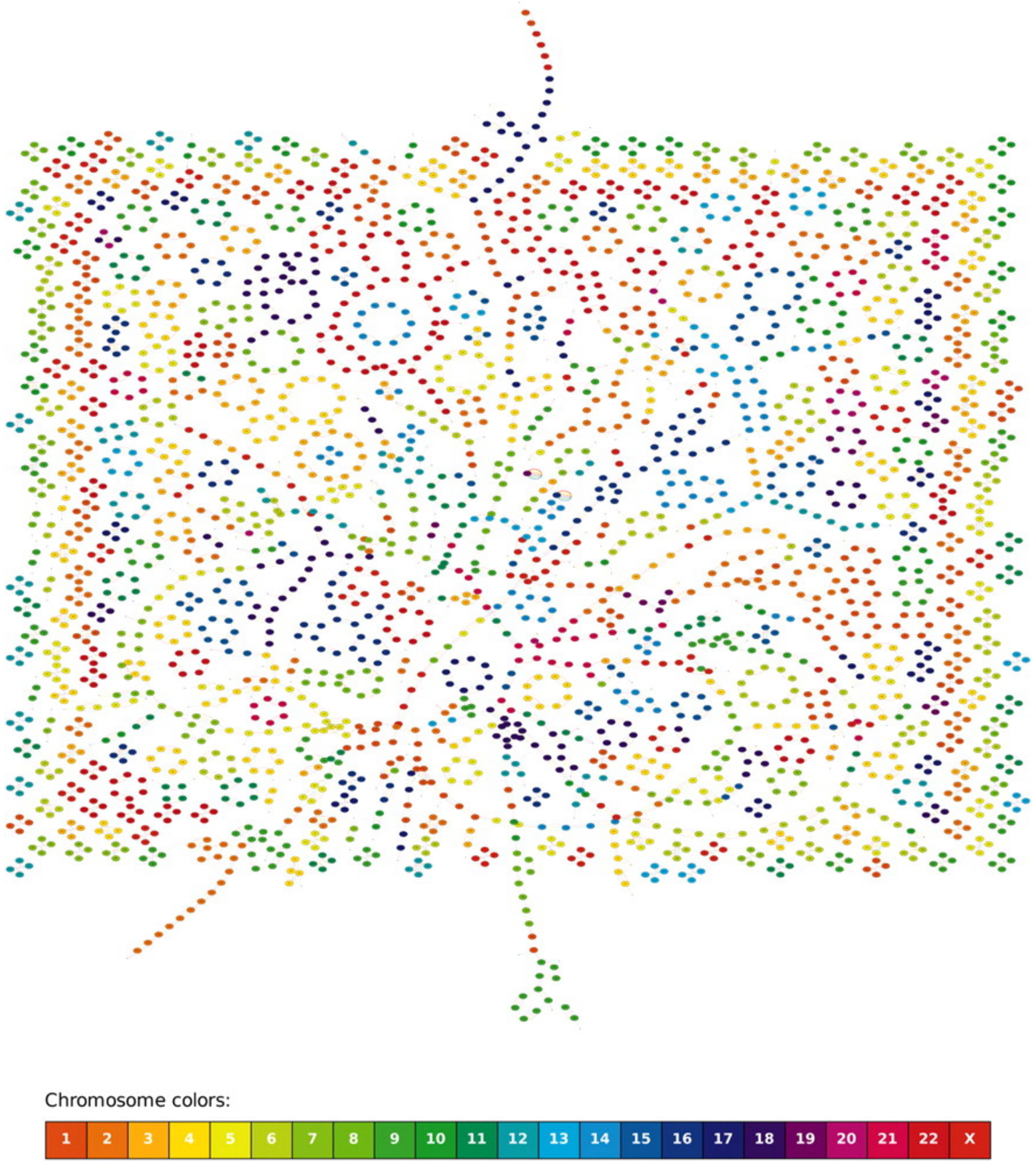
http://en.wikipedia.org/wiki/Buffalo_buffalo_Buffalo_buffalo_buffalo_buffalo_Buffalo_buffalo



New York bison whom other New York bison bully, themselves bully New York bison.

http://en.wikipedia.org/wiki/Buffalo_buffalo_Buffalo_buffalo_buffalo_buffalo_Buffalo_buffalo

BUFFALO BUFFALO OF VISUALIZATION



Alekseyev, M.A. and P.A. Pevzner, Breakpoint graphs and ancestral genome reconstructions. Genome Res, 2009. 19(5): p. 943-57.

SATISFY YOUR AUDIENCE, NOT YOURSELF.

Be aware of bias in evaluating effectiveness of visual forms.

BMJ

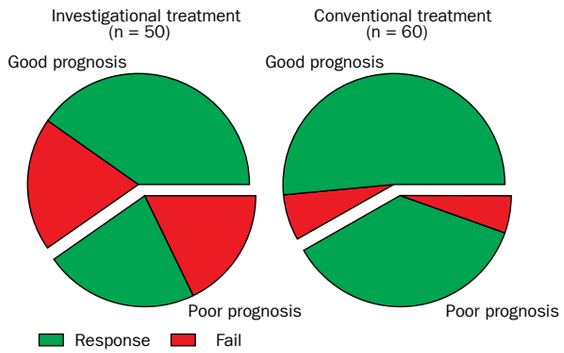
Influence of data display formats on physician investigators' decisions to stop clinical trials: prospective trial with repeated measures

Linda S Elting, Charles G Martin, Scott B Cantor, Edward B Rubenstein

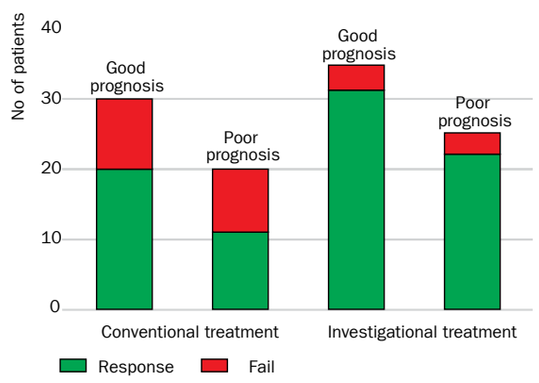
table

	Conventional treatment		Investigational treatment	
	Total no	% Fail	Total no	% Fail
Good prognosis	30	30	35	11
Poor prognosis	20	45	25	12
Total	50	38	60	12

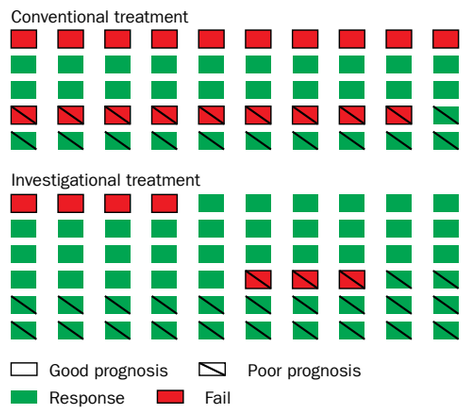
pie chart



bar graph



icon graph

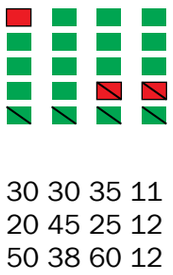
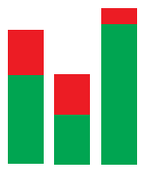


Elting et al. BMJ 316 (1999).

BMJ

Influence of data display formats on physician investigators' decisions to stop clinical trials: prospective trial with repeated measures

Linda S Elting, Charles G Martin, Scott B Cantor, Edward B Rubenstein

	% accuracy	% preference
 30 30 35 11 20 45 25 12 50 38 60 12	82	0
	68	62
	56	24
	56	14

Elting et al. BMJ 316 (1999).

BMJ

Influence of data display formats on physician investigators' decisions to stop clinical trials: prospective trial with repeated measures

Linda S Elting, Charles G Martin, Scott B Cantor, Edward B Rubenstein

"...eight voiced considerable contempt for the [icon] display."

Elting et al. BMJ 316 (1999).

BMJ

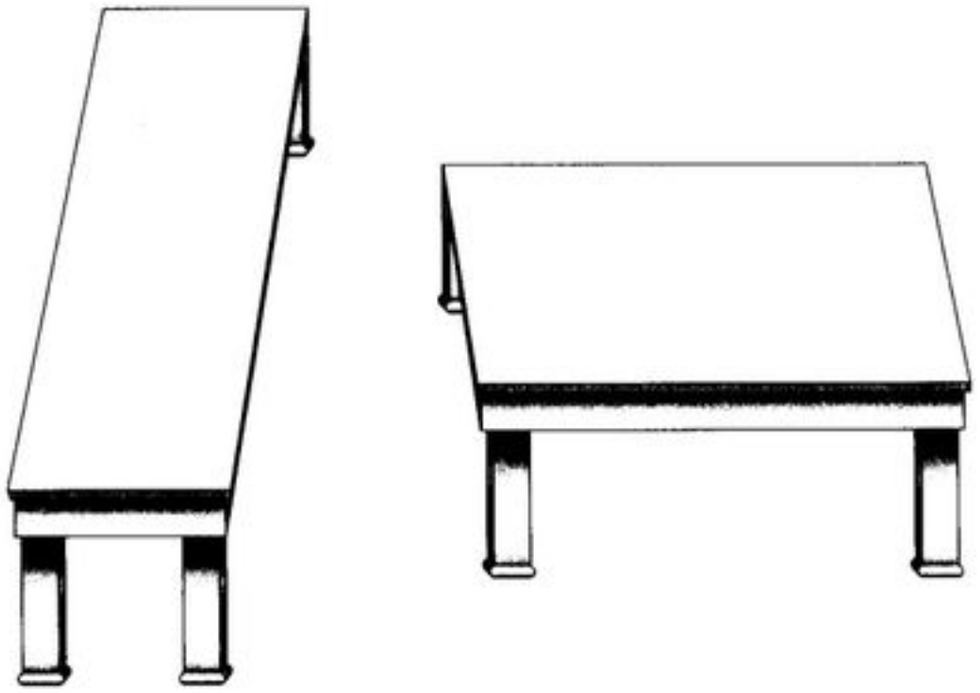
Influence of data display formats on physician investigators' decisions to stop clinical trials: prospective trial with repeated measures

Linda S Elting, Charles G Martin, Scott B Cantor, Edward B Rubenstein

"... icon displays were often preferred by nurses, students, ... but were considered unacceptable by physicians."

Elting et al. BMJ 316 (1999).

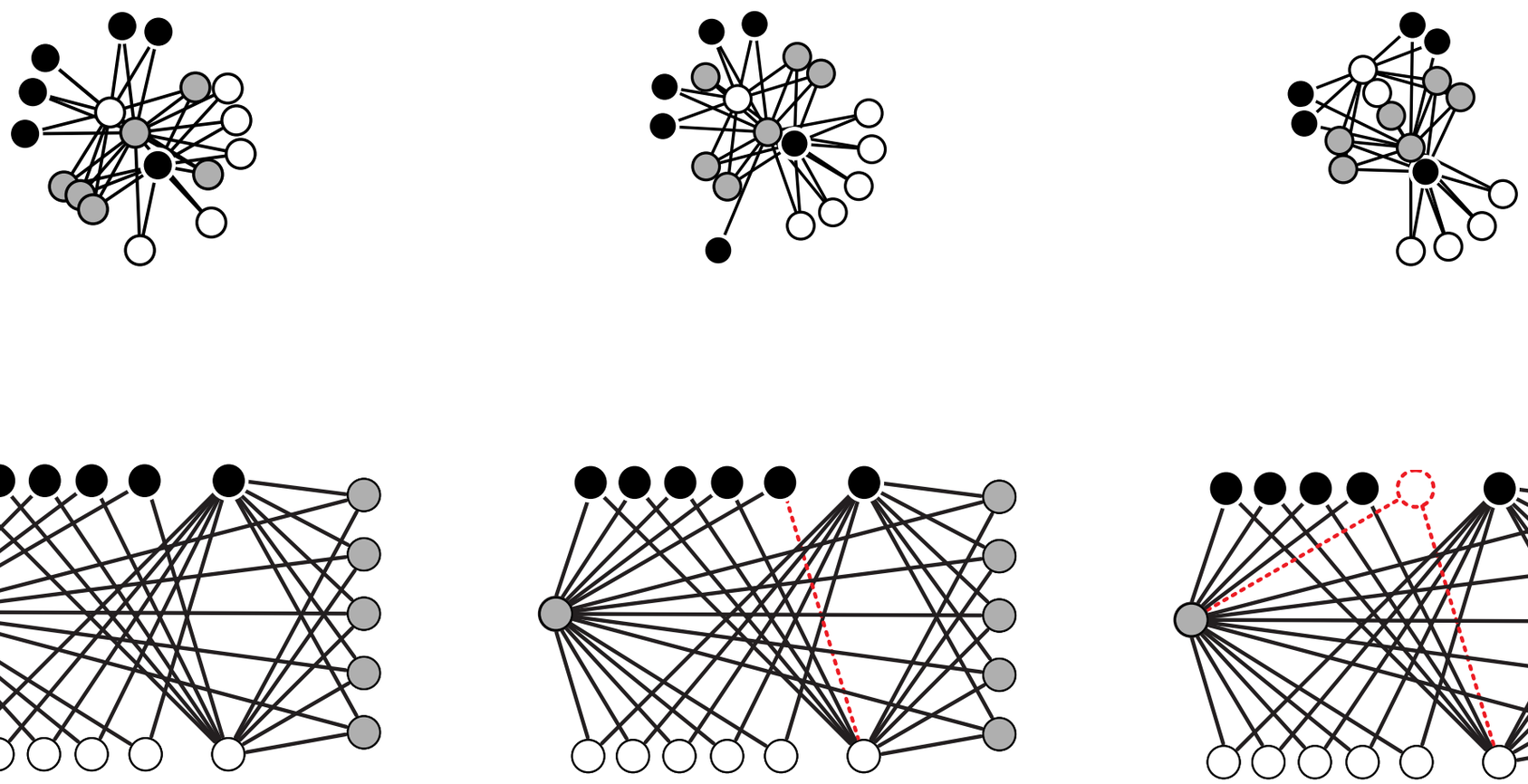
WE ARE EASILY DECEIVED



Shephard's tables, http://www.michaelbach.de/ot/sze_shepardTables/

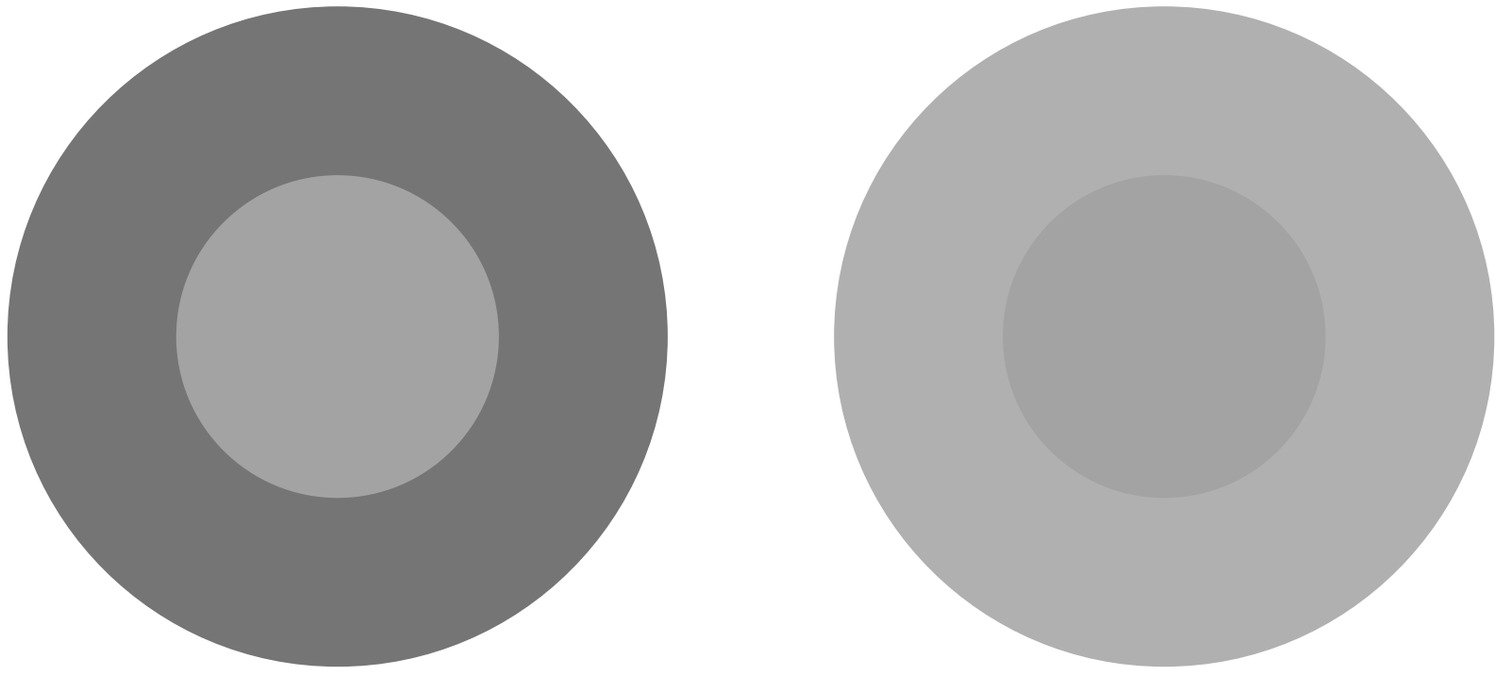
DO NOT BE CHARMED BY INEFFECTIVE FORMS

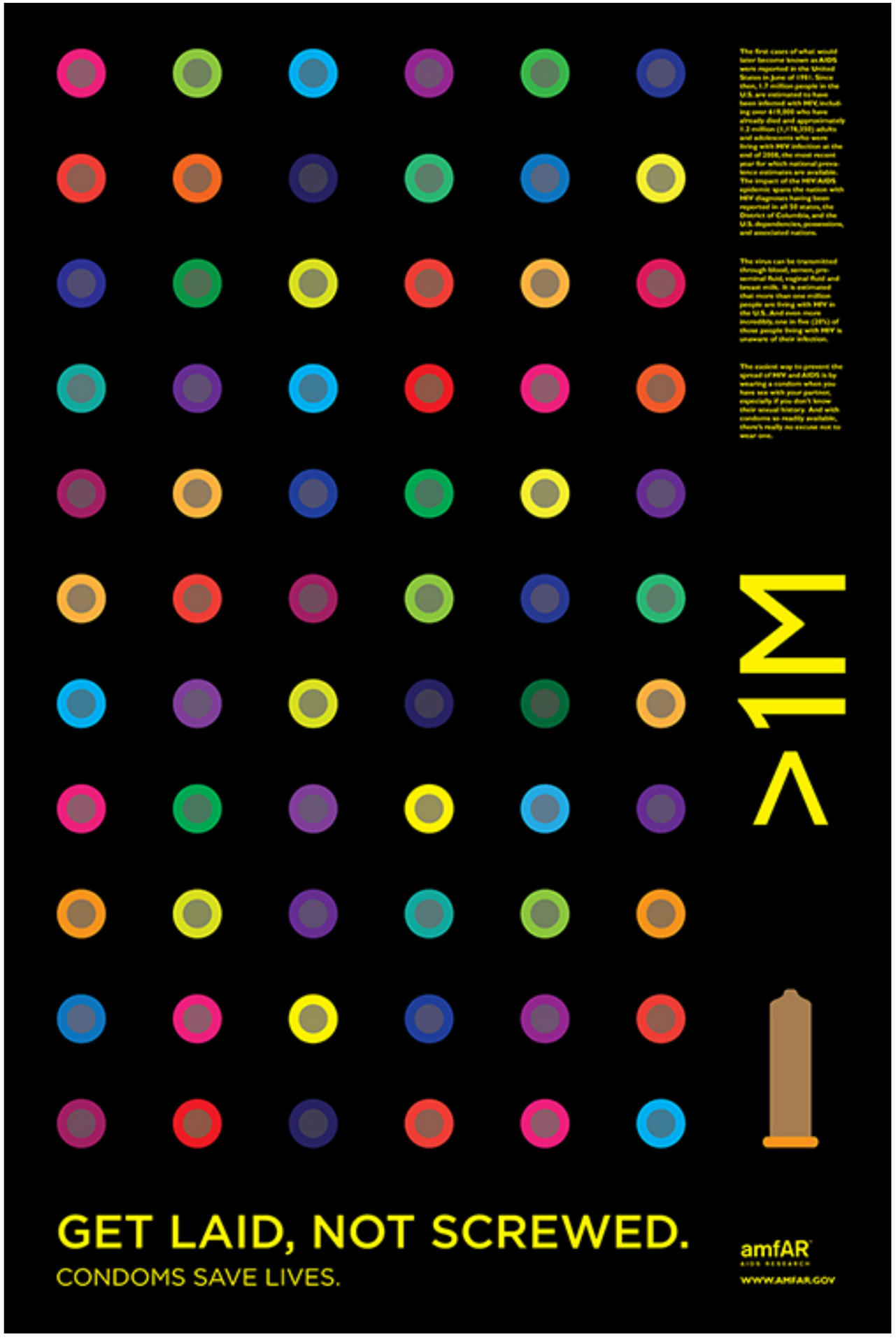
THE INCOMPARABLE NETWORK



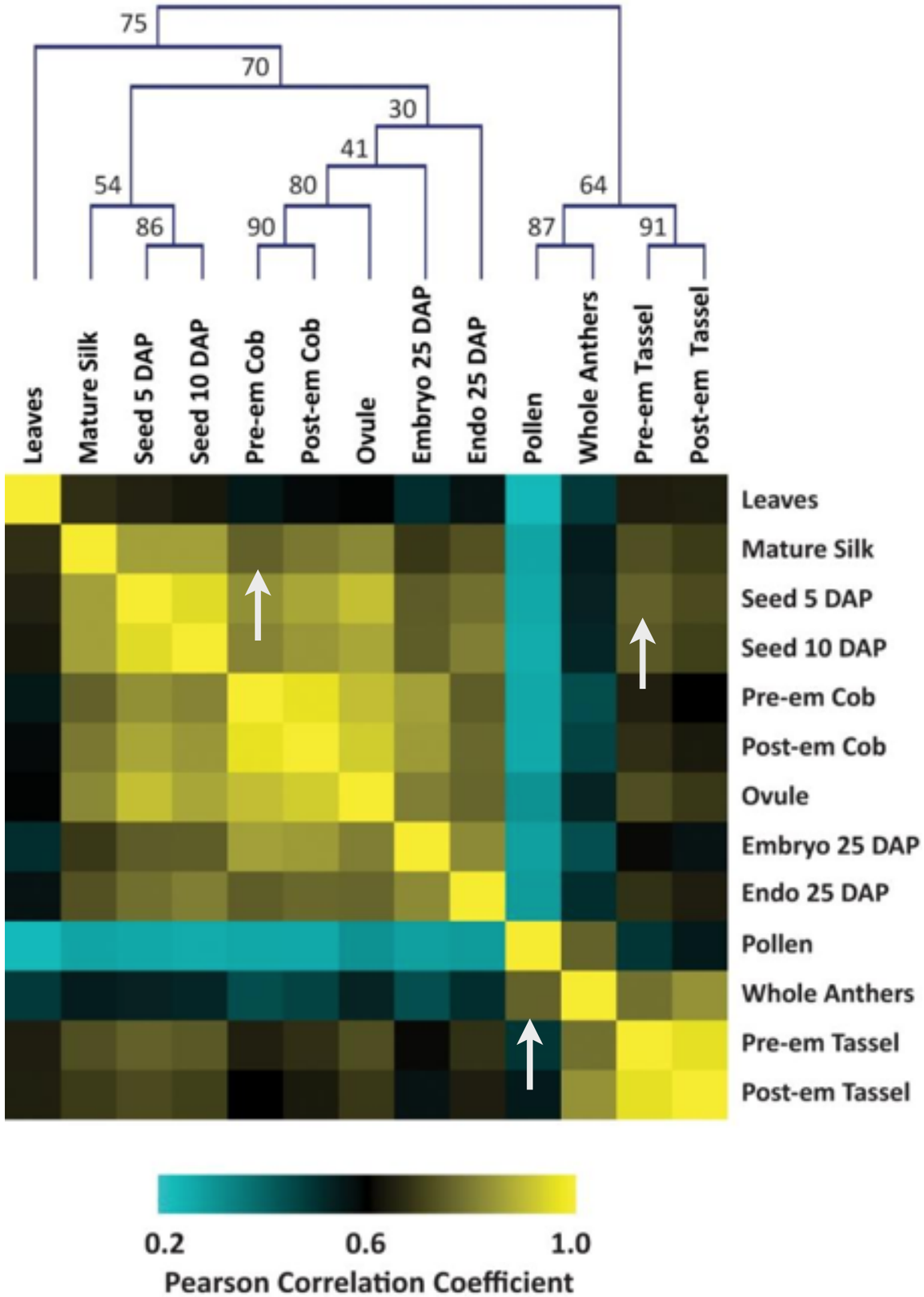
Krzywinski et al. Briefings in Bioinformatics (2011).

RESPECT HUMAN VISUAL LIMITATIONS

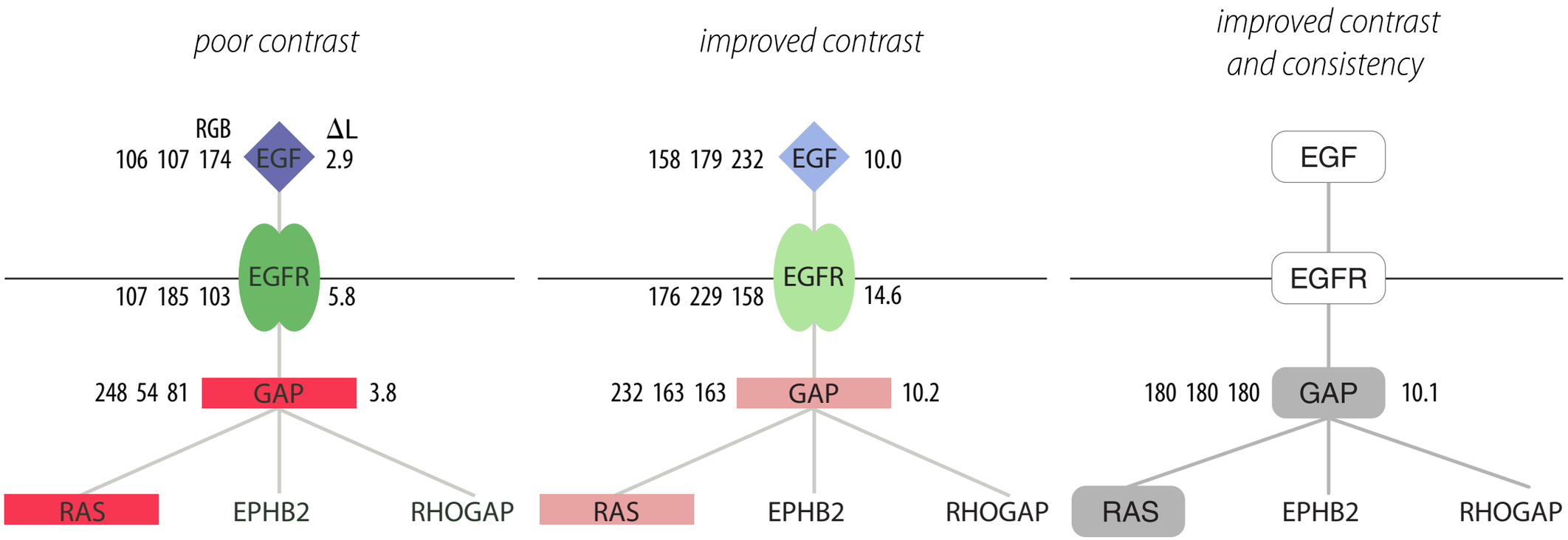




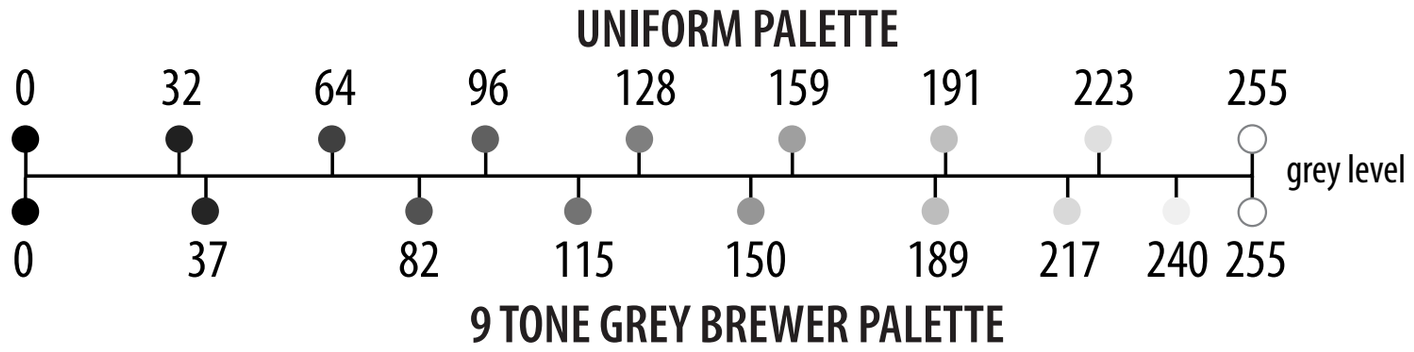
http://elenamiska.com/AIDS-Poster

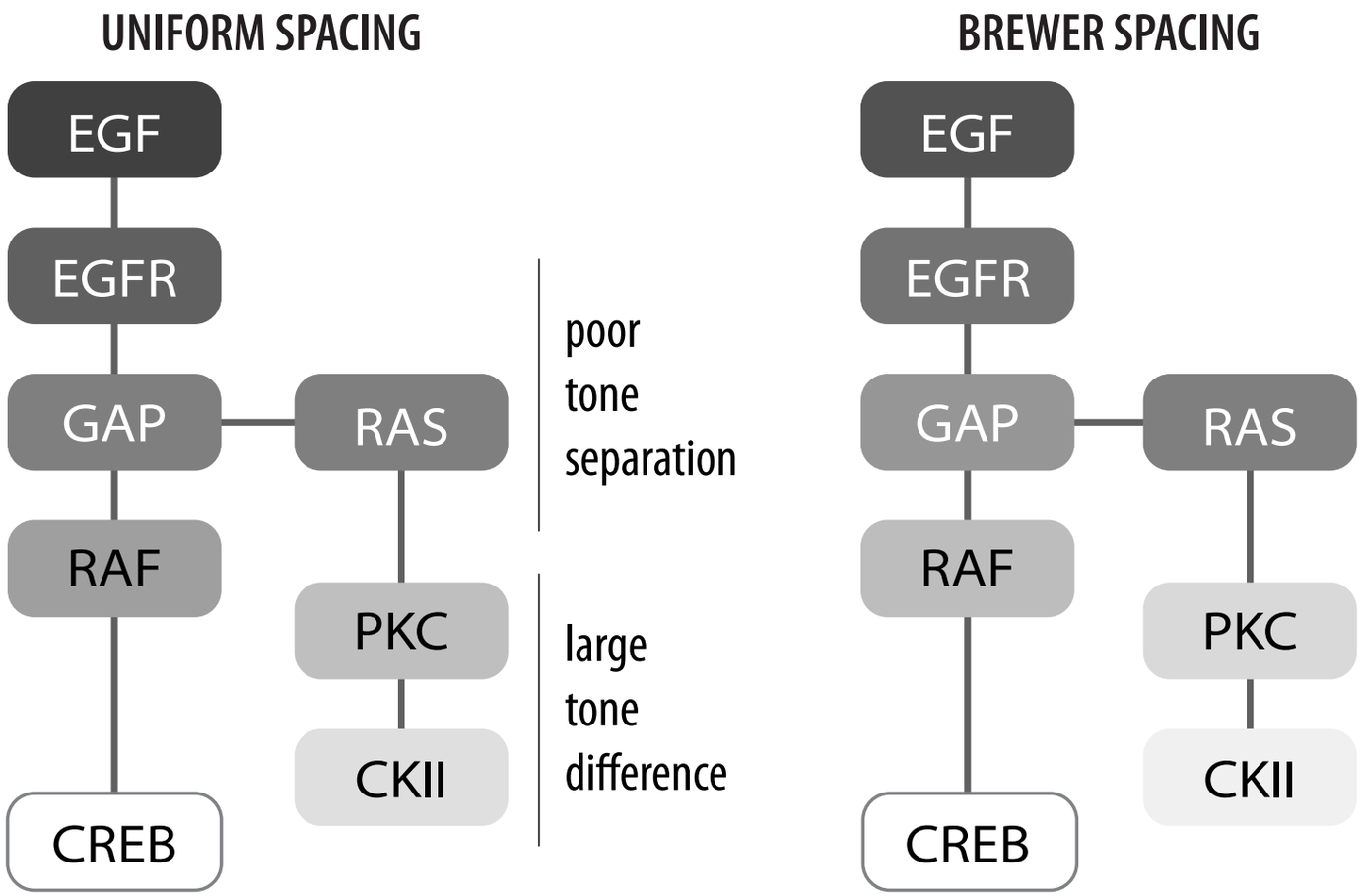


Utility of RNA Sequencing for Analysis of Maize Reproductive Transcriptomes Plant Gen. 2011 4: 191–203 10.3835/plantgenome2011.05.0015



Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, et al. (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Res 13: 2363–2371.

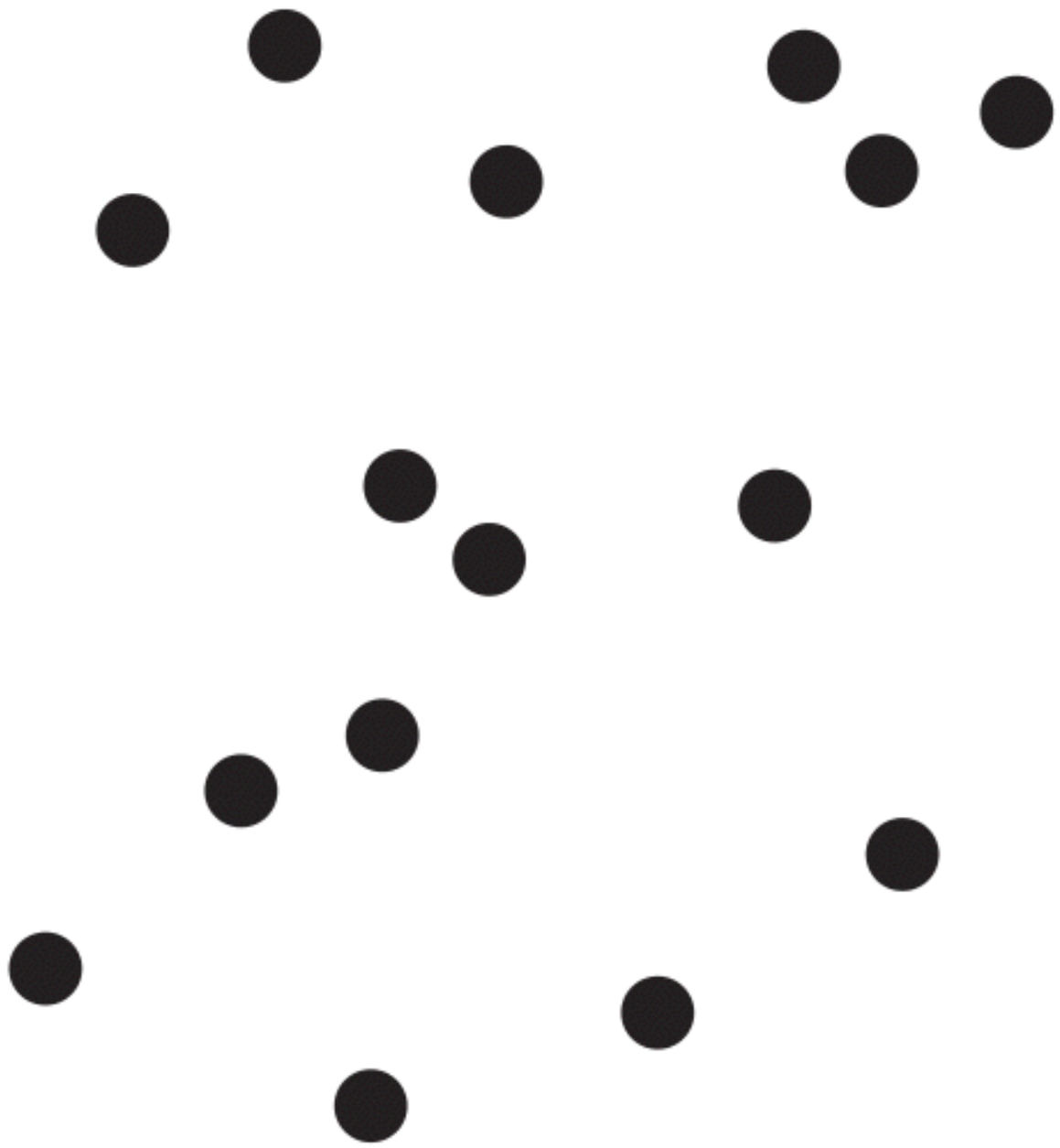

UNIFORM PALETTE
0
32
64
96
128
159
191
223
255
grey level
0
37
82
115
150
189
217
240
255
9 TONE GREY BREWER PALETTE


UNIFORM SPACING
BREWER SPACING
EGF
EGFR
GAP
RAS
RAF
PKC
CKII
CREB
poor
tone
separation
large
tone
difference

DATA INFORMS VARIATION

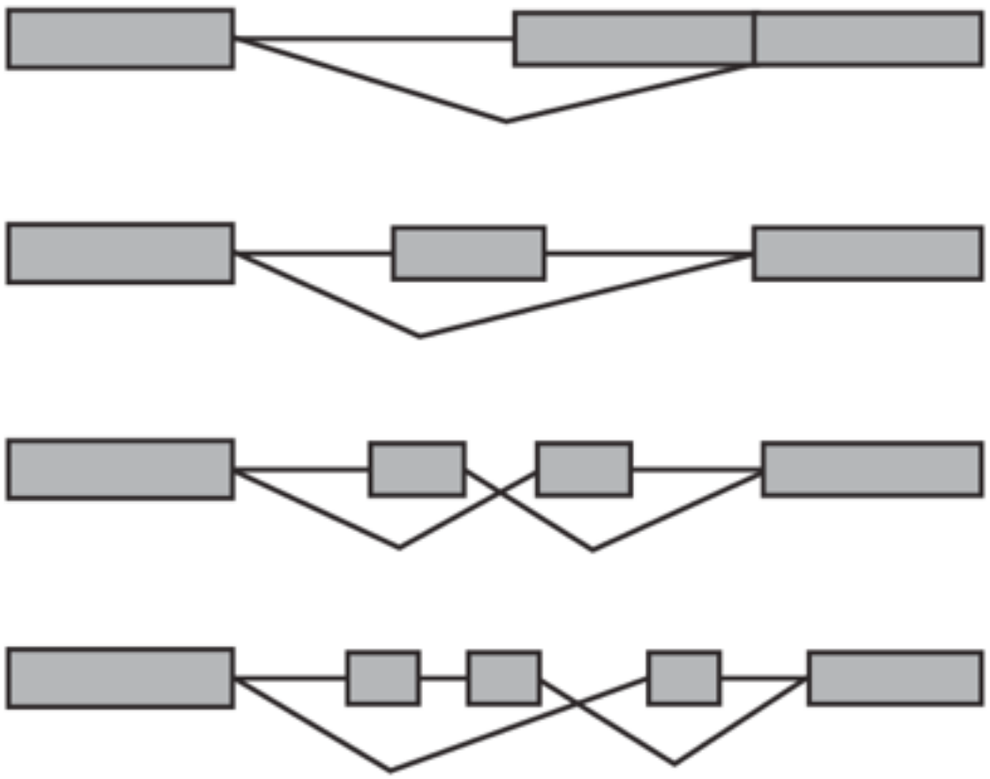
Patterns are hard to see when variation is due to both data and formatting.

HOW MANY DOTS?

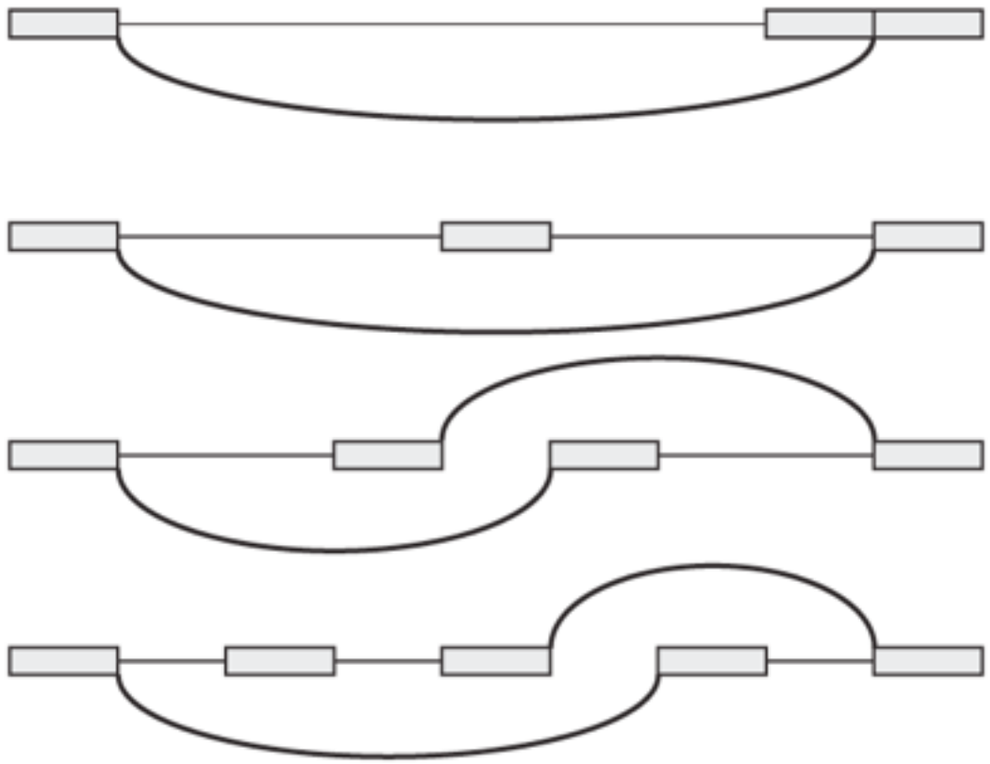


AVOID EXCESS DEGREES OF FREEDOM

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.

CONSISTENCY

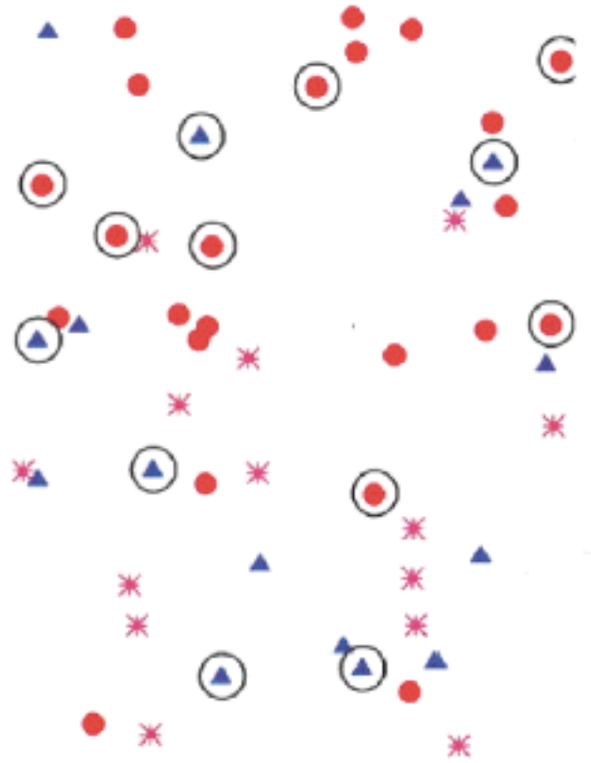
Avoid the use of similar encodings for independent variables.

Consistency

- Visual variation in a figure should always reflect and enhance any underlying variation in the data.
- Avoid using more than one encoding to communicate the same information.
- Do not use visually similar encodings for independent variables

Consistency

- red - processed genes, but salience attenuated
  - other genes encoded with competing glyph - red star.



confusing encoding

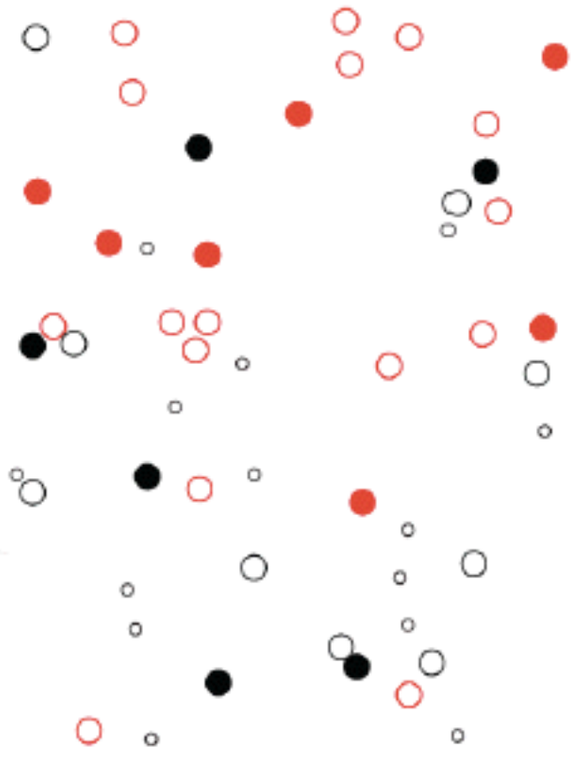
PS

NPS

Gene

transcribed

(doesn't appear in legend)



hierarchical encoding

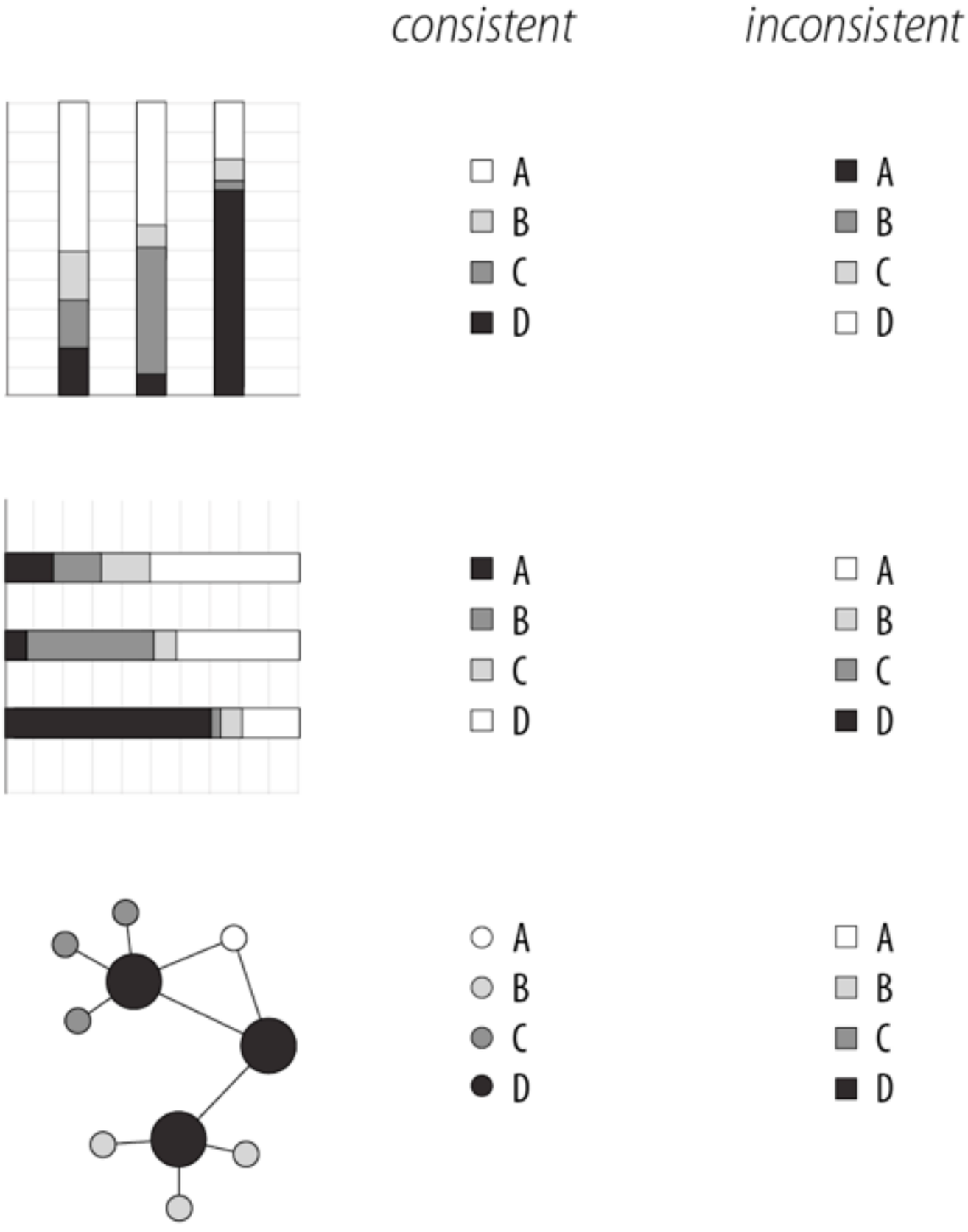
pseudogenes

processed \ transcribed	N	Y
N	○	●
Y	○	●

other genes ○

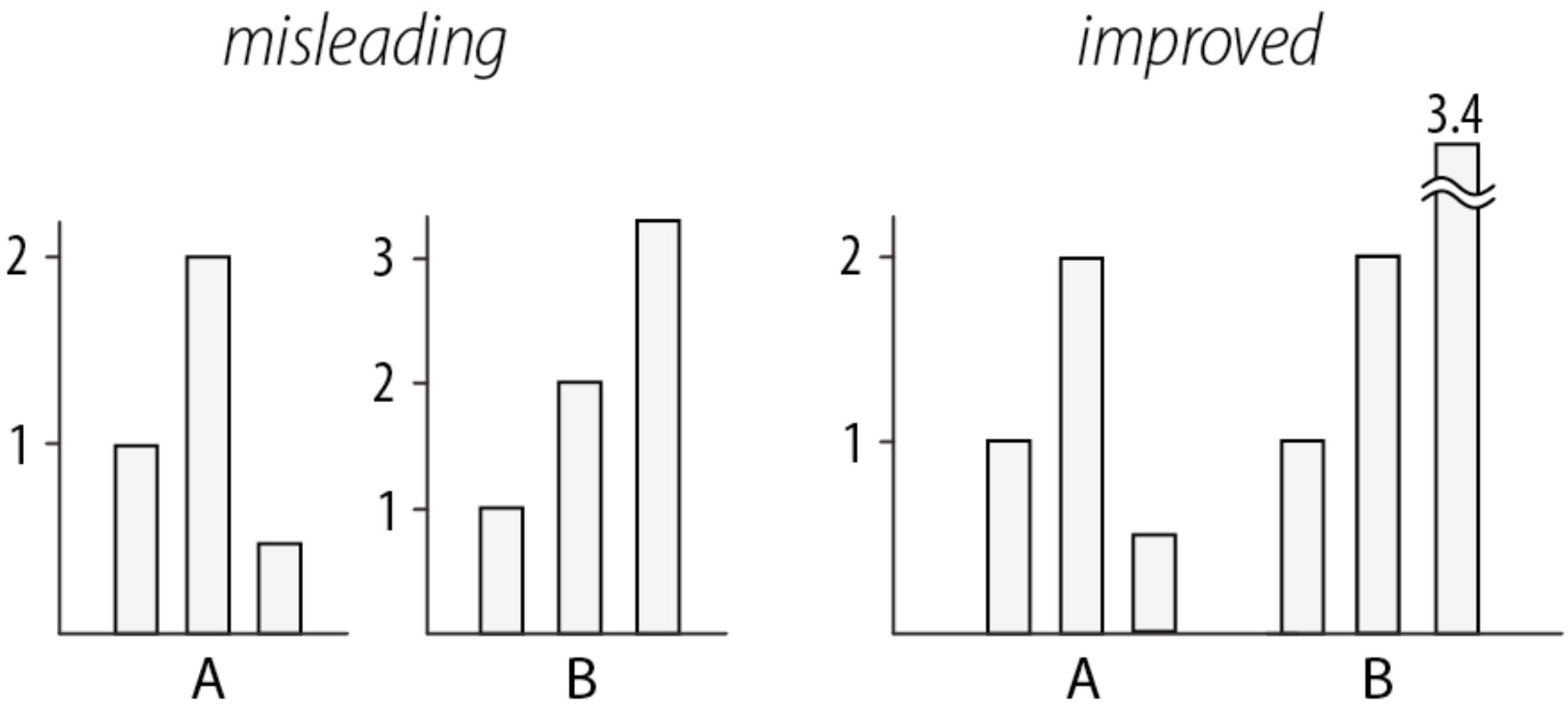
Consistency

- Order items in a legend according to order of appearance in the plot

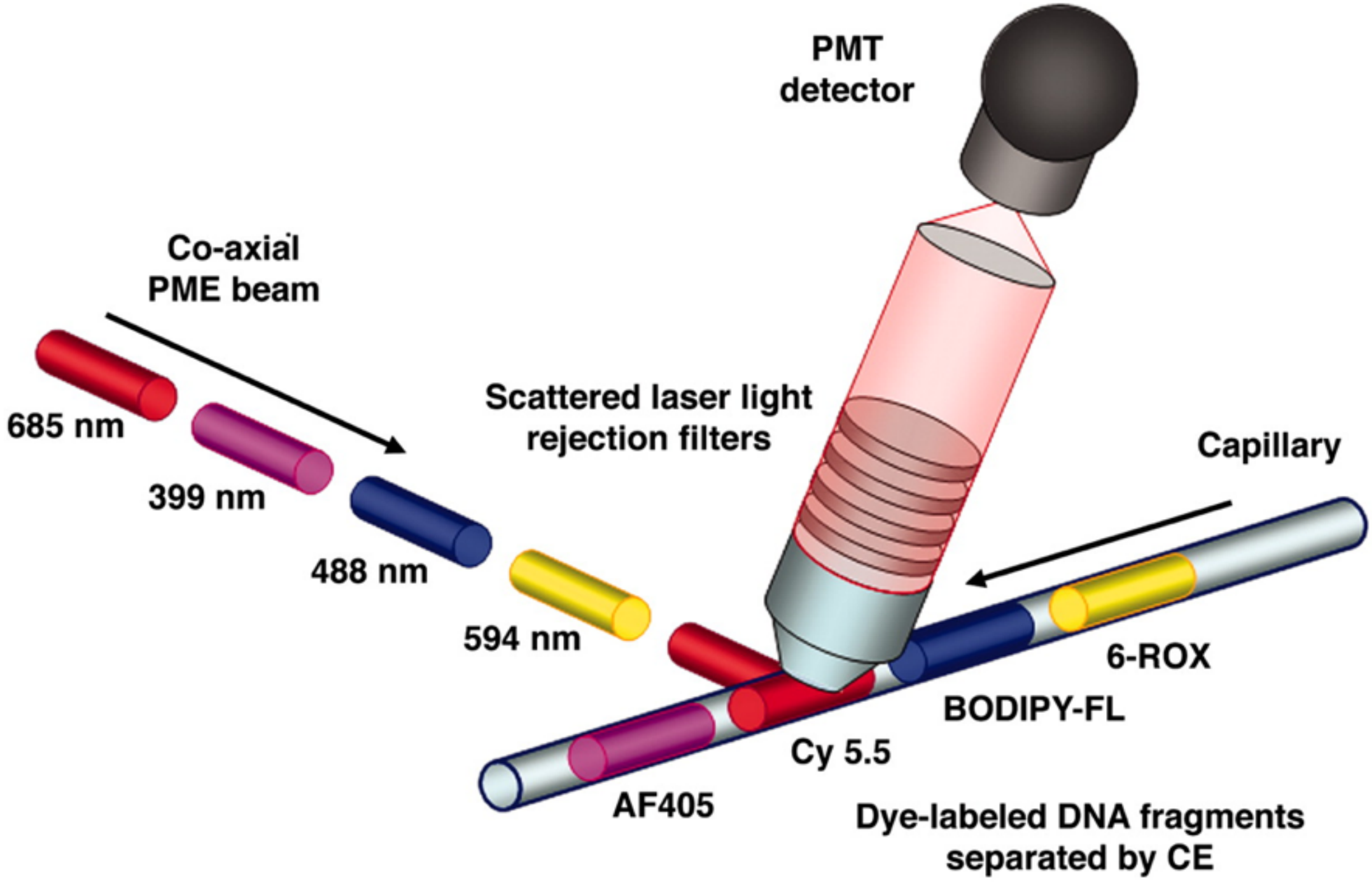


Consistency - Navigational aids

- Use consistent axes when comparing charts



e.g. Fig 1 in Raina SZ, Faith JJ, Disotell TR, Seligmann H, Stewart CB, et al. (2005) Evolution of base-substitution gradients in primate mitochondrial genomes. Genome Res 15: 665-673.



Lewis EK, Haaland WC, Nguyen F, Heller DA, Allen MJ, et al. (2005) Color-blind fluorescence detection for four-color DNA sequencing. Proc Natl Acad Sci U S A 102: 5346-5351.

homonymous encoding

685 nm CY 5.5

399 nm AF405

488 nm BODIPY-FL

594 nm 6-ROX

disambiguated

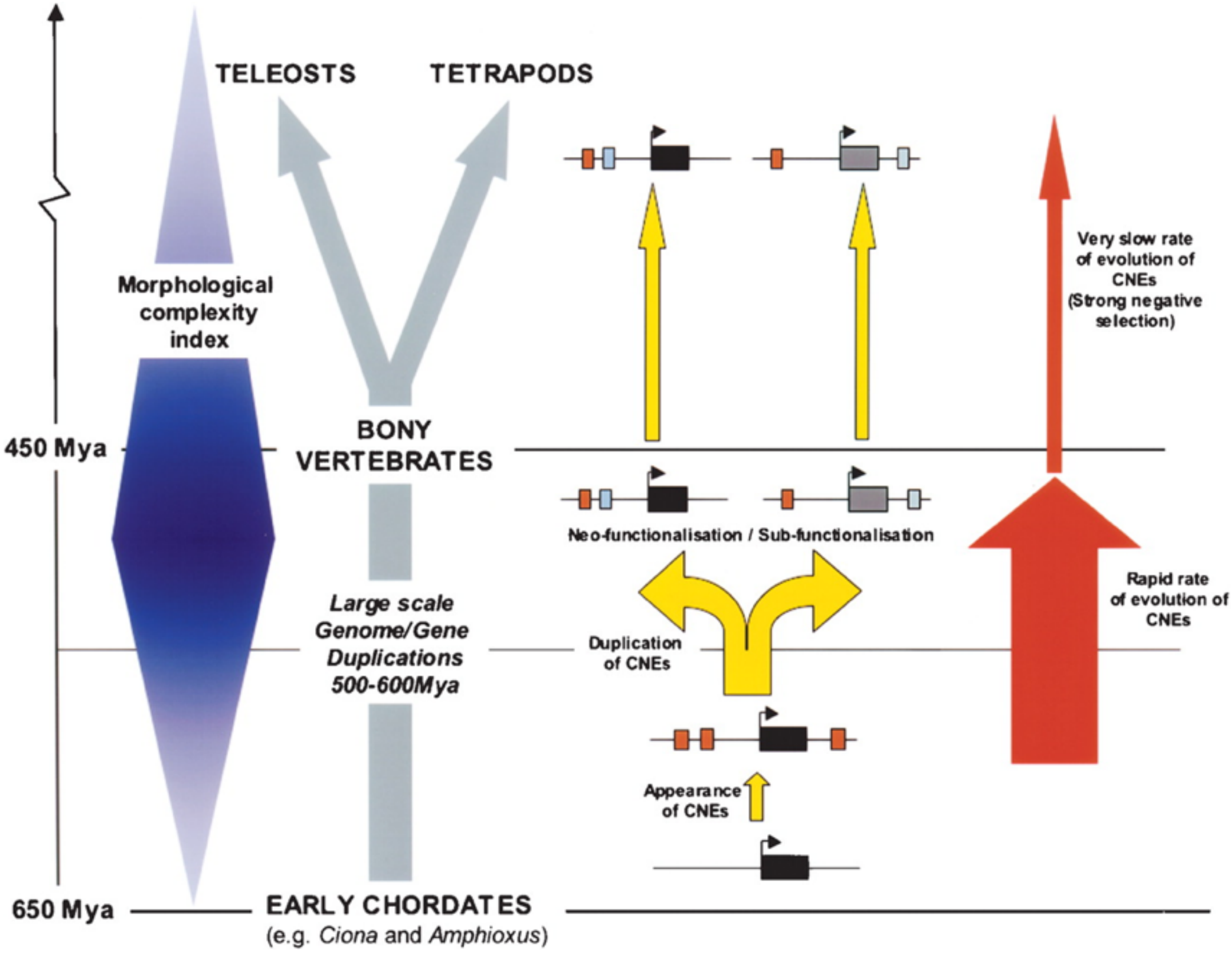
685 nm T CY 5.5

399 nm C AF405

488 nm A BODIPY-FL

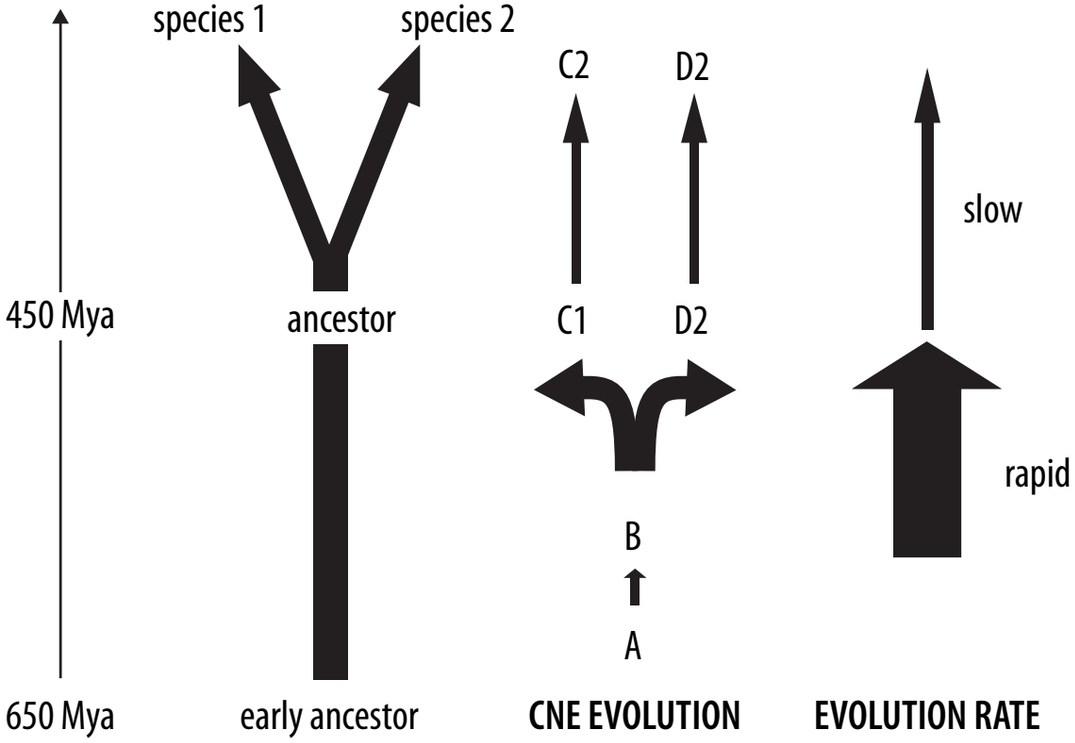
594 nm G 6-ROX

Lewis EK, Haaland WC, Nguyen F, Heller DA, Allen MJ, et al. (2005) Color-blind fluorescence detection for four-color DNA sequencing. Proc Natl Acad Sci U S A 102: 5346-5351.

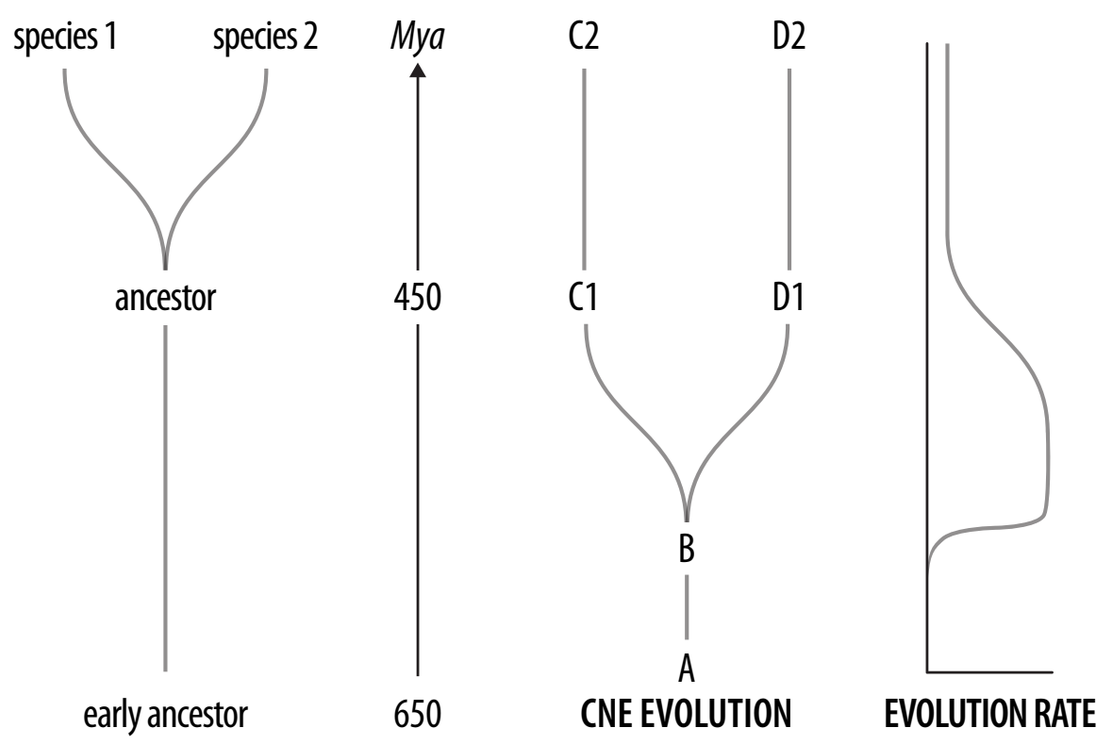


McEwen GK, Woolfe A, Goode D, Vavouri T, Callaway H, et al. (2006) Ancient duplicated conserved noncoding elements in vertebrates: a genomic and functional analysis. Genome Res 16: 451–465.

arrows too complex



simplified arrows



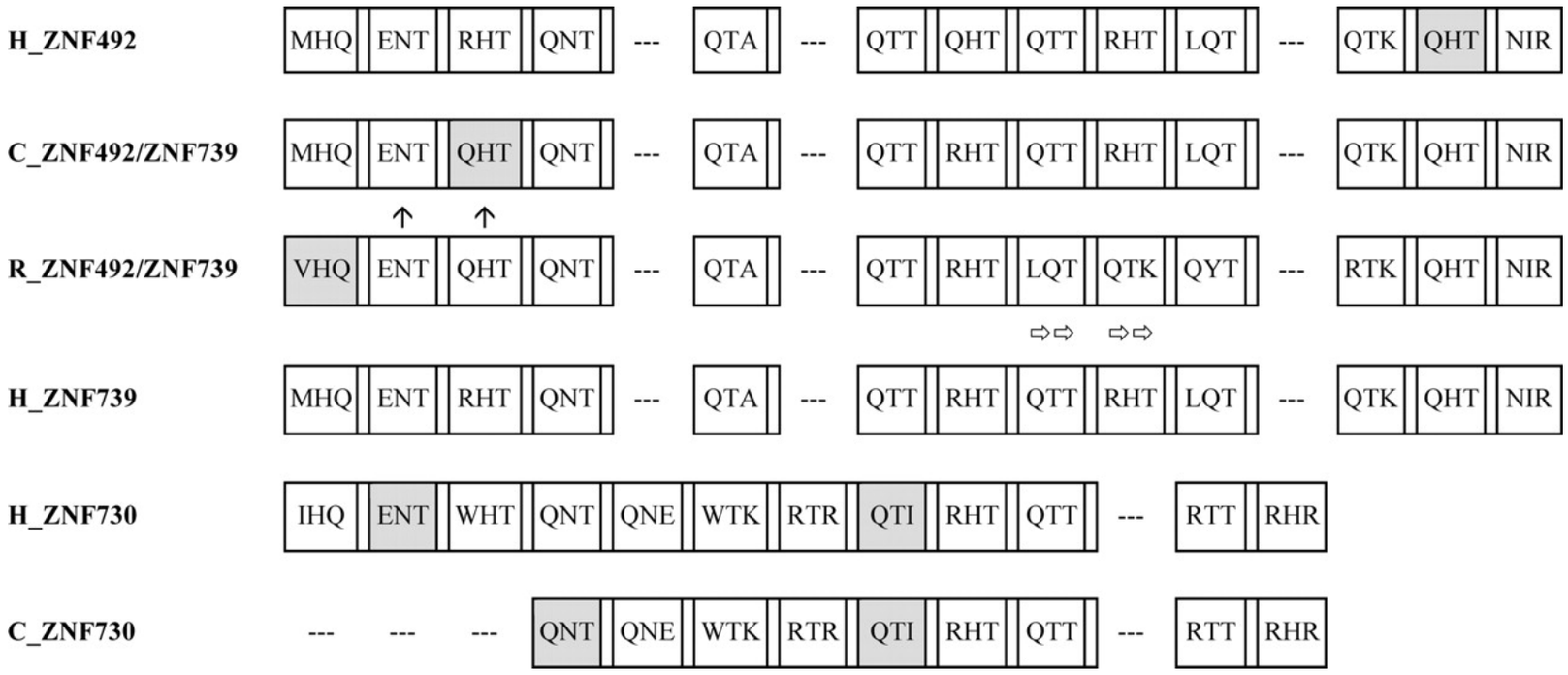
McEwen GK, Woolfe A, Goode D, Vavouri T, Callaway H, et al. (2006) Ancient duplicated conserved noncoding elements in vertebrates: a genomic and functional analysis. Genome Res 16: 451–465.

Consistency - numbers

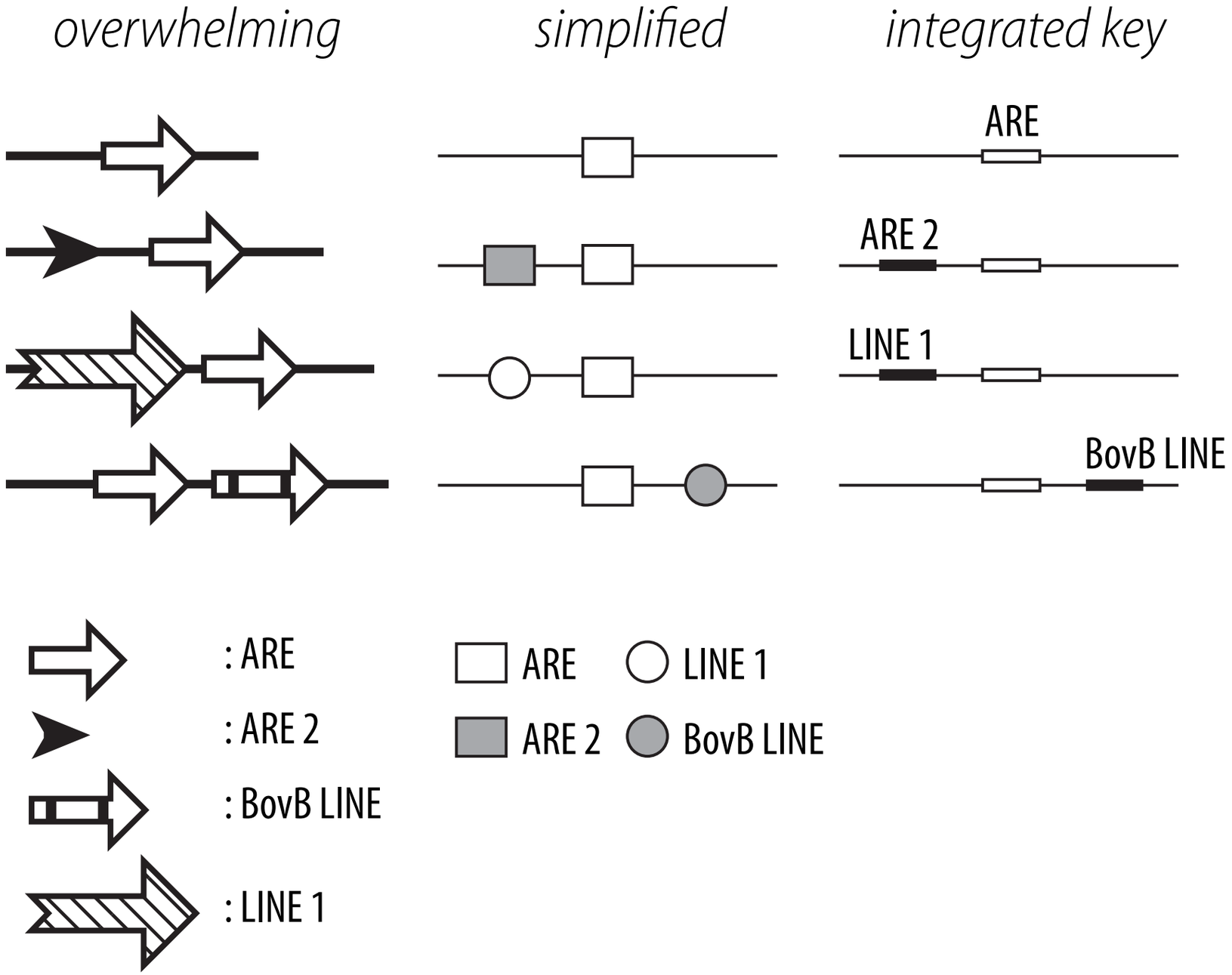
- Use the same amount of digits in the numbers - (avoid e.g. 234000, 34.567 together)
- Percentage should always sum up to 100%

REDUNDANCY

Don’t repeat yourself. Don’t repeat yourself.



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.



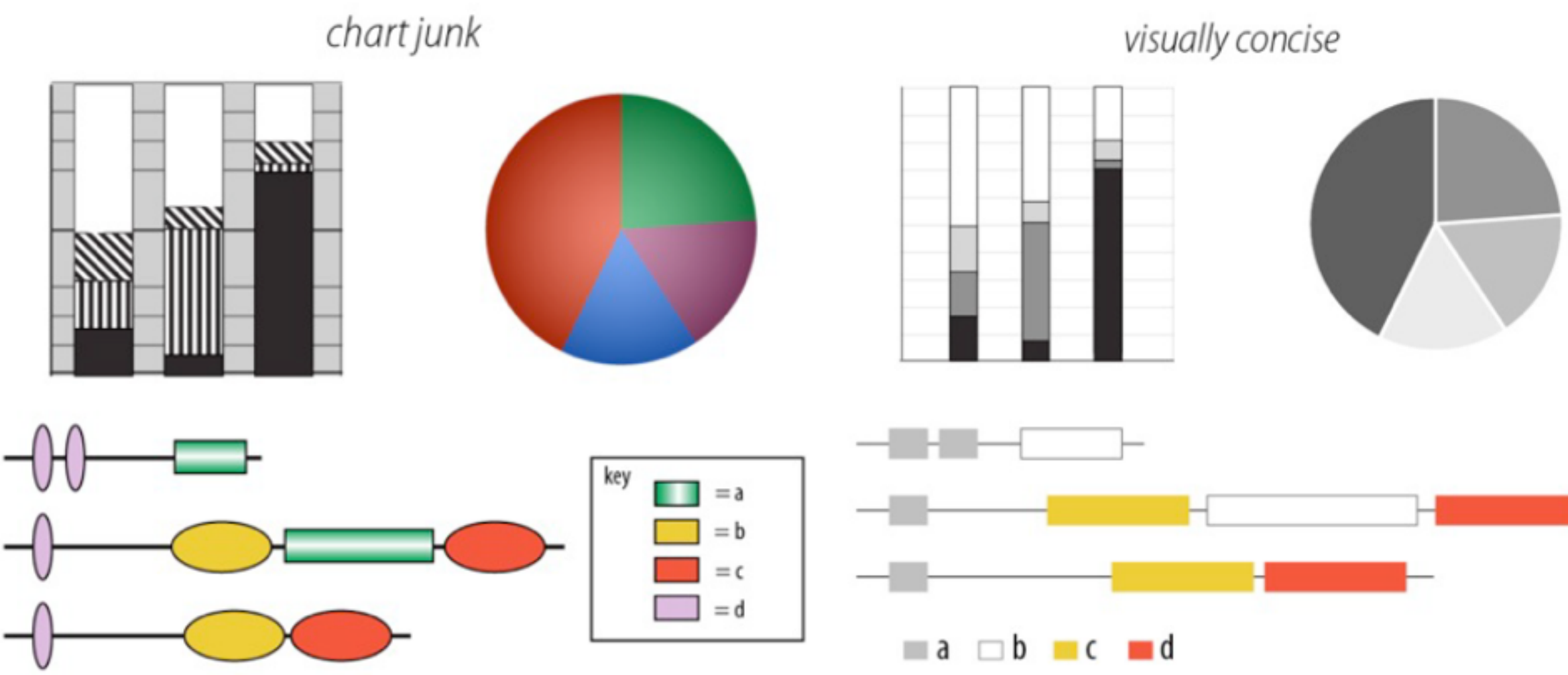
Nikaido M, Rooney AP, Okada N (1999) Phylogenetic relationships among cetartiodactyls based on insertions of short and long interspersed elements: Hippopotamuses are the closest extant relatives of whales. Proceedings of the National Academy of Sciences 96: 10261–10266.

CONCISENESS

Use the fewest elements possible—keep data-to-ink ratio high

Increase data:ink ratio

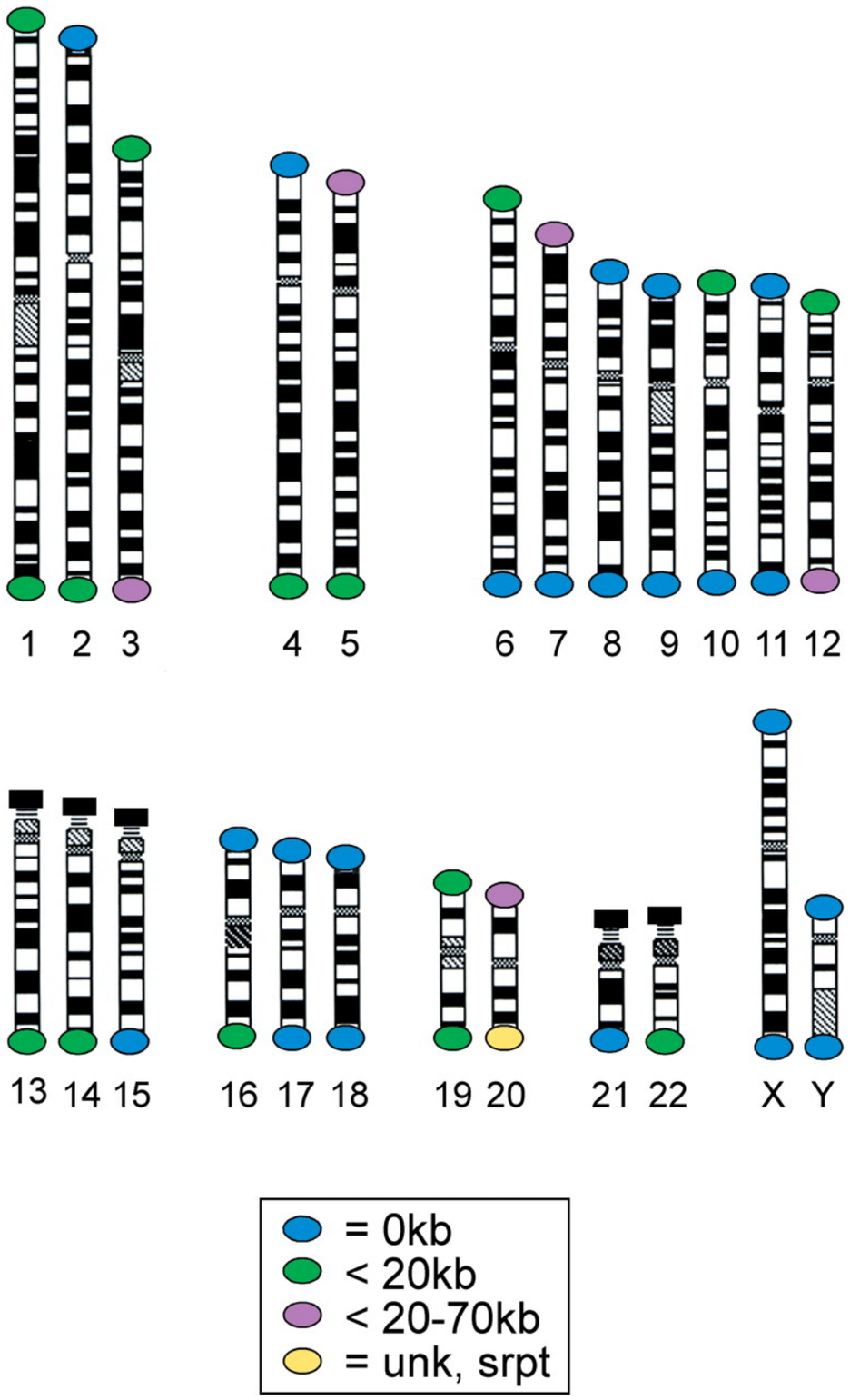
- Avoid "Chart junk"



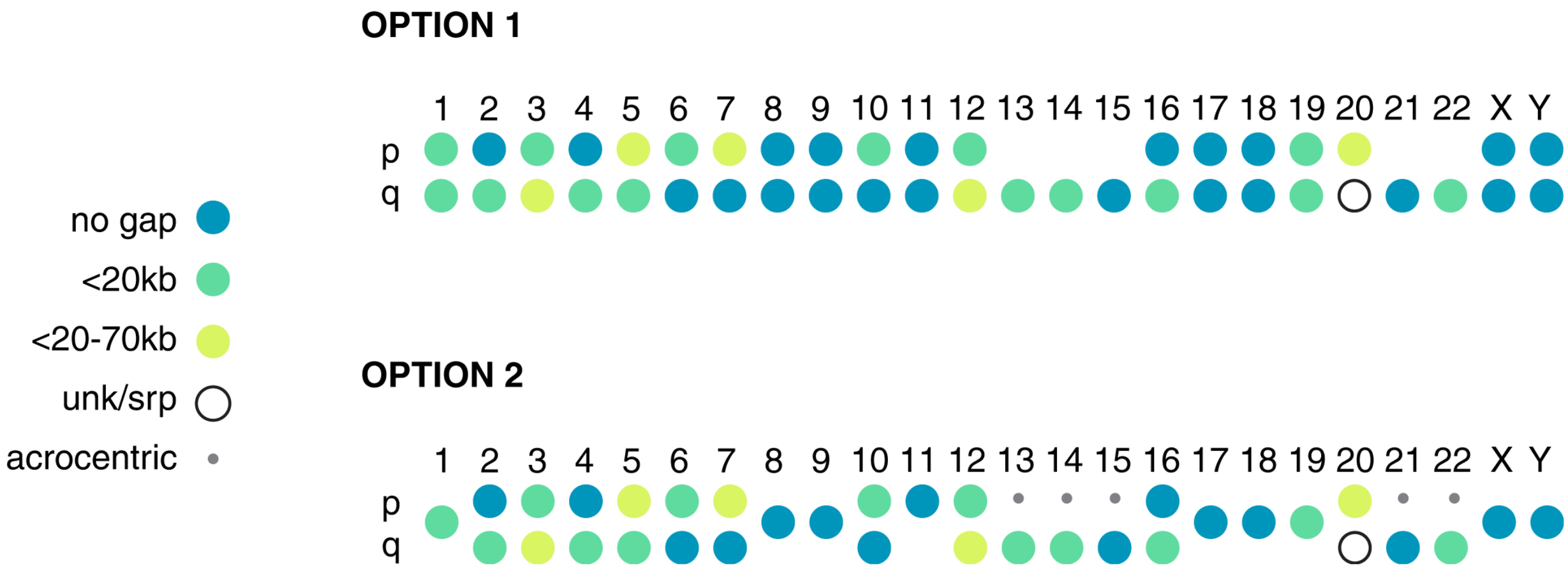
REMOVE TO IMPROVE

Use the fewest elements possible—keep data-to-ink ratio high.

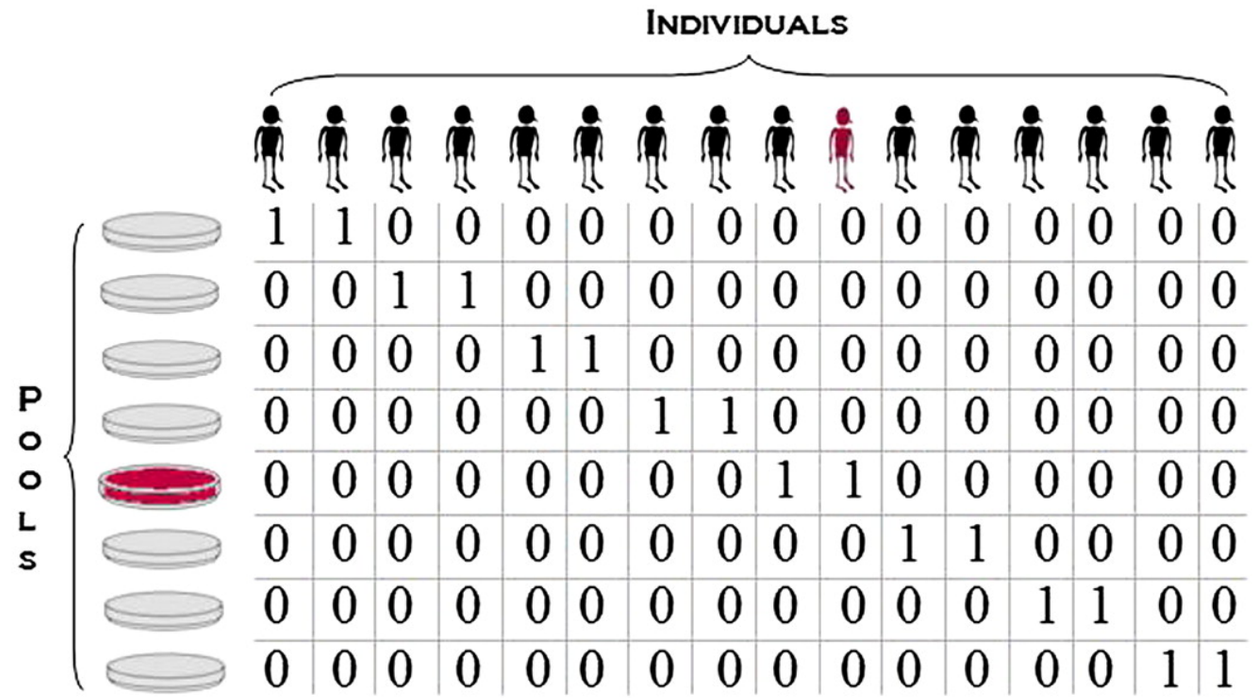
Shelter your reader from unnecessary complexity.



Telomere sequence gaps. Riethman, H., et al., Mapping and initial analysis of human subtelomeric sequence assemblies. Genome Res, 2004. 14(1): p. 18-28.



Telomere sequence gaps. Riethman, H., et al., Mapping and initial analysis of human subtelomeric sequence assemblies. Genome Res, 2004. 14(1): p. 18-28.
redesign by M Krzywinski



Resequencing with naïve and log pool designs. Prabhu, S. and I. Pe'er, Overlapping pools for high-throughput targeted resequencing. Genome Res, 2009. 19(7): p. 1254-61.

busy

DISJOINT

INDIVIDUALS

POOLS

●	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
●	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0
●	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0
●	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0
●	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
●	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0
●	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0
●	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

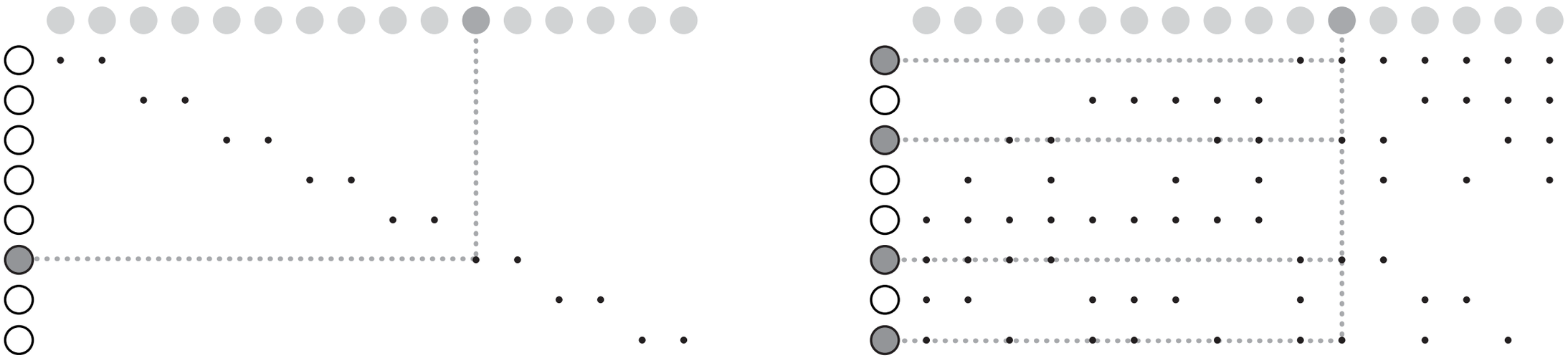
OVERLAPPING

INDIVIDUALS

POOLS

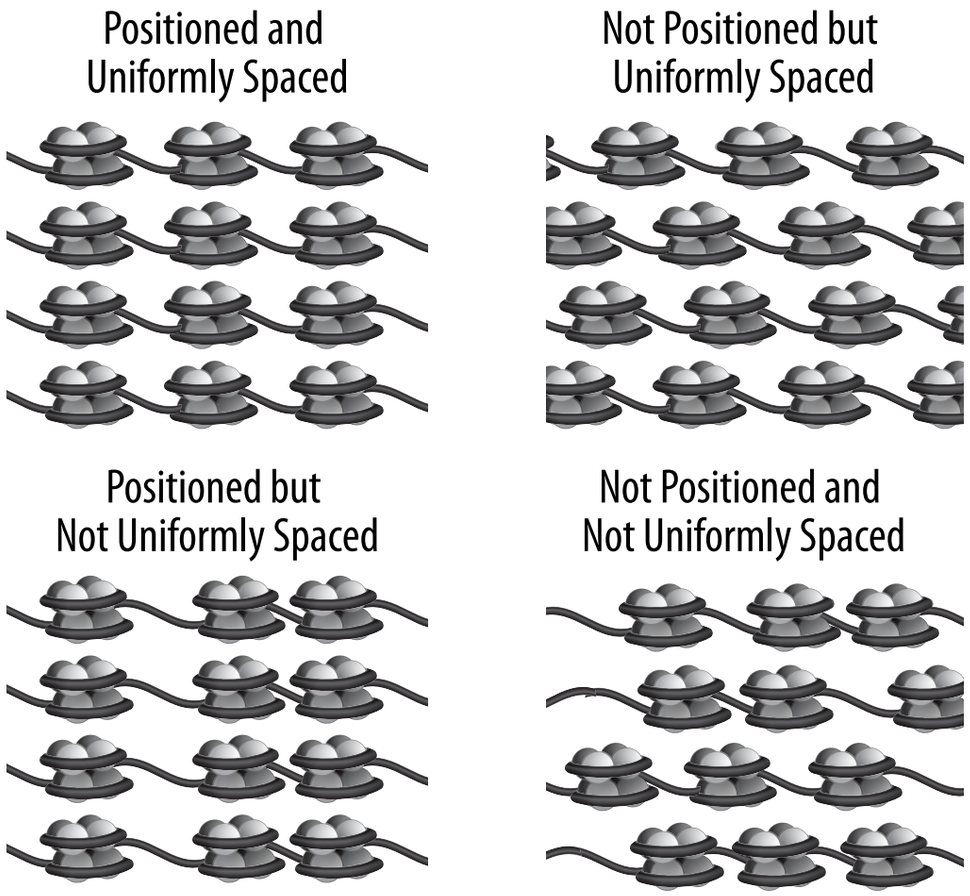
●	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1
●	0	0	0	0	1	1	1	1	0	0	0	0	1	1	1
●	0	0	1	1	0	0	1	1	0	0	1	1	0	0	0
●	0	1	0	1	0	1	0	1	0	1	0	1	0	0	1
●	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0
●	1	1	1	1	0	0	0	0	1	1	1	1	0	0	0
●	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0
●	1	0	1	0	1	0	1	0	1	0	1	0	1	0	0

improved

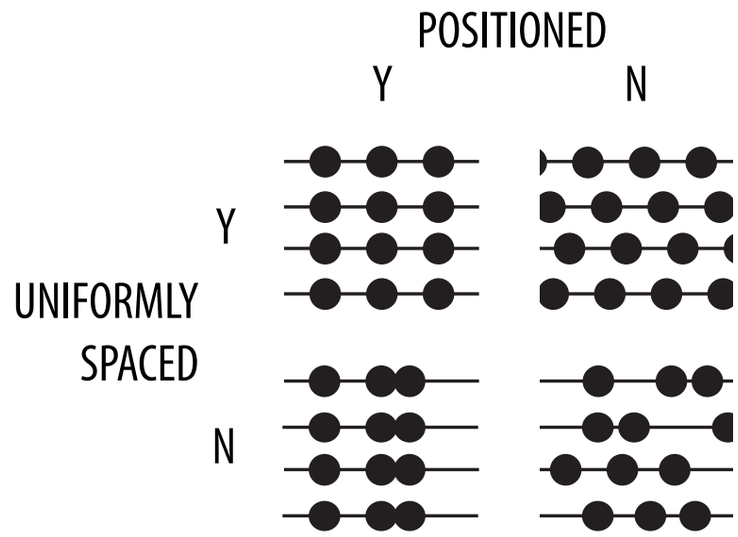


Resequencing with naïve and log pool designs. Prabhu, S. and I. Pe'er, Overlapping pools for high-throughput targeted resequencing. Genome Res, 2009. 19(7): p. 1254-61.
redesign by M Krzywinski

detail exposed



detail encapsulated



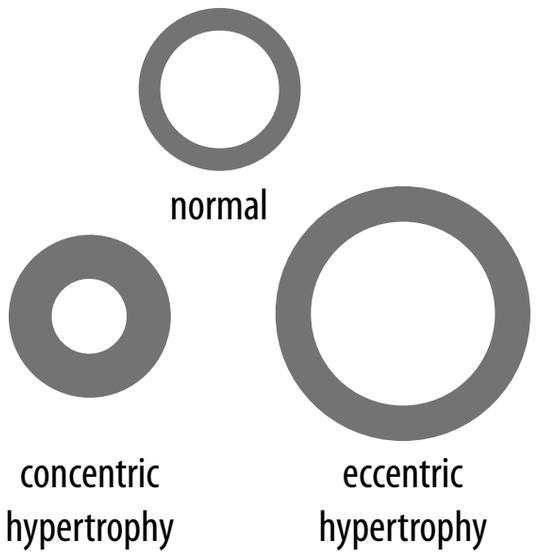
Valouev A, Ichikawa J, Tonthat T, Stuart J, Ranade S, et al. (2008) A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. Genome Res 18: 1051–1063.

CLARITY

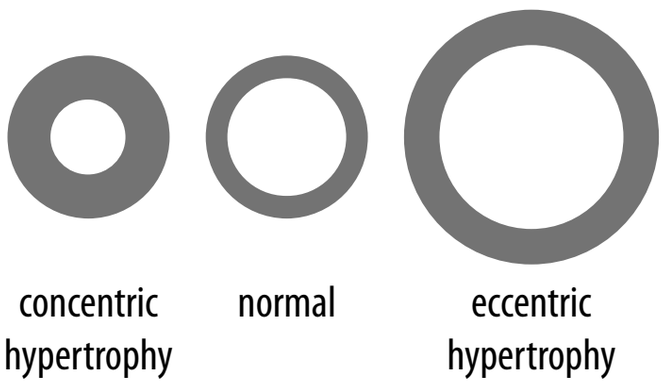
Make sure that elements are visible and unobscured.

Don’t count on your audience to figure out what you mean. Say it.

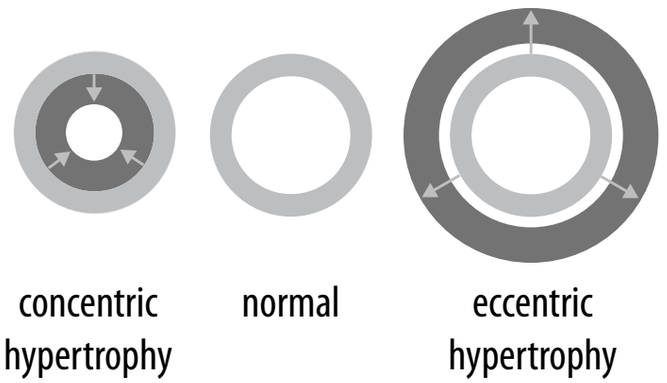
perceptually challenging



helpful, but not explicit



helpful and explicit



Nadeau JH, Burrage LC, Restivo J, Pao YH, Churchill G, et al. (2003) Pleiotropy, homeostasis, and functional networks based on assays of cardiovascular traits in genetically randomized populations. Genome Res 13: 2082–2091.

FOCUS & EMPHASIS

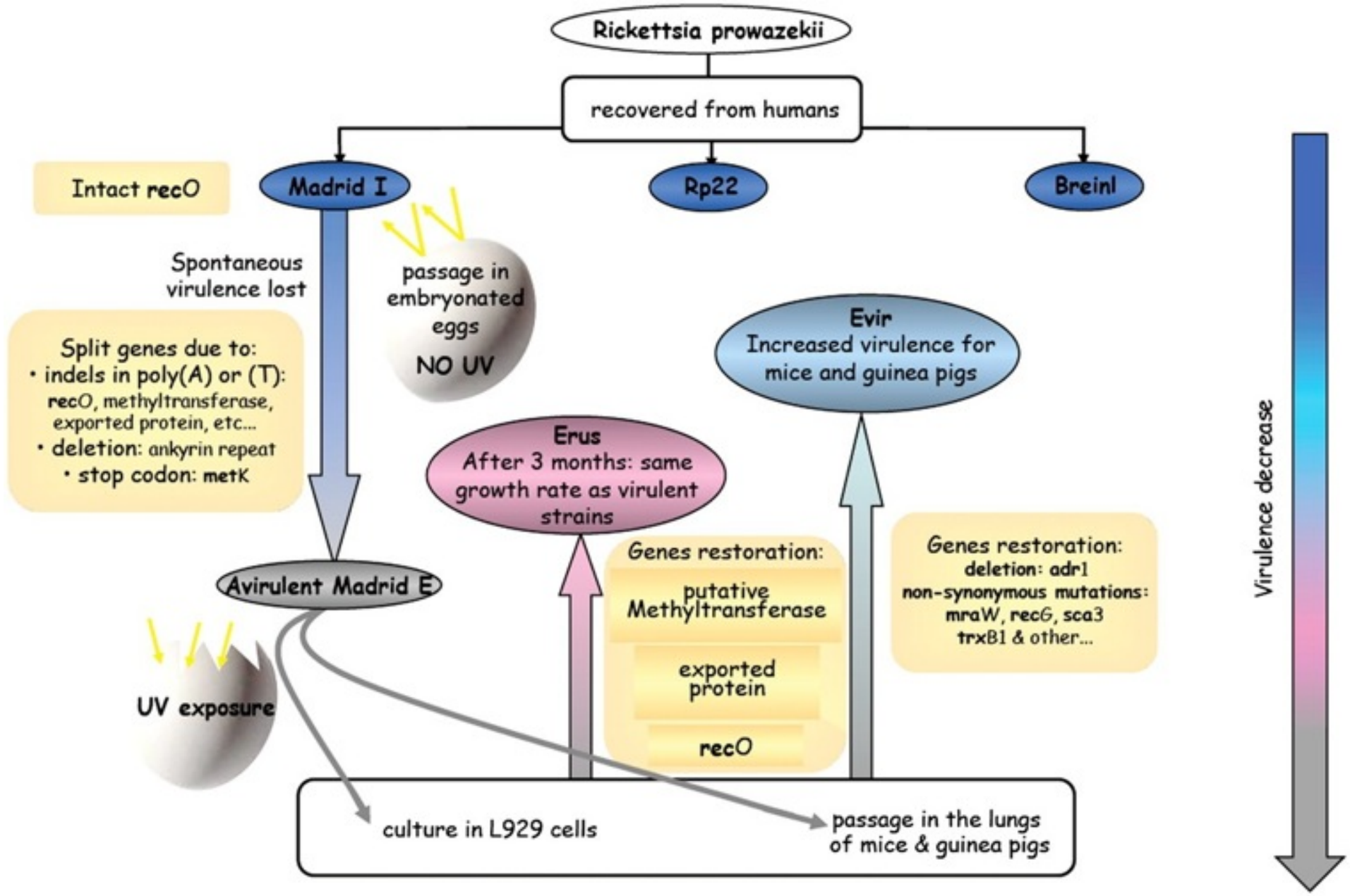
Match the pertinence of an object with its visual salience.

Apply visual organization Gestalt principles.

Marcin Ryczek, http://plfoto.com/182312

Marcin Ryczek, http://plfoto.com/182312

EVERYTHING IS EMPHASIZED



Bechah, Y., et al., Genomic, proteomic, and transcriptomic analysis of virulent and avirulent Rickettsia prowazekii reveals its adaptive mutation capabilities. Genome Res, 2010. 20(5): p. 655-63.

NOTHING IS EMPHASIZED

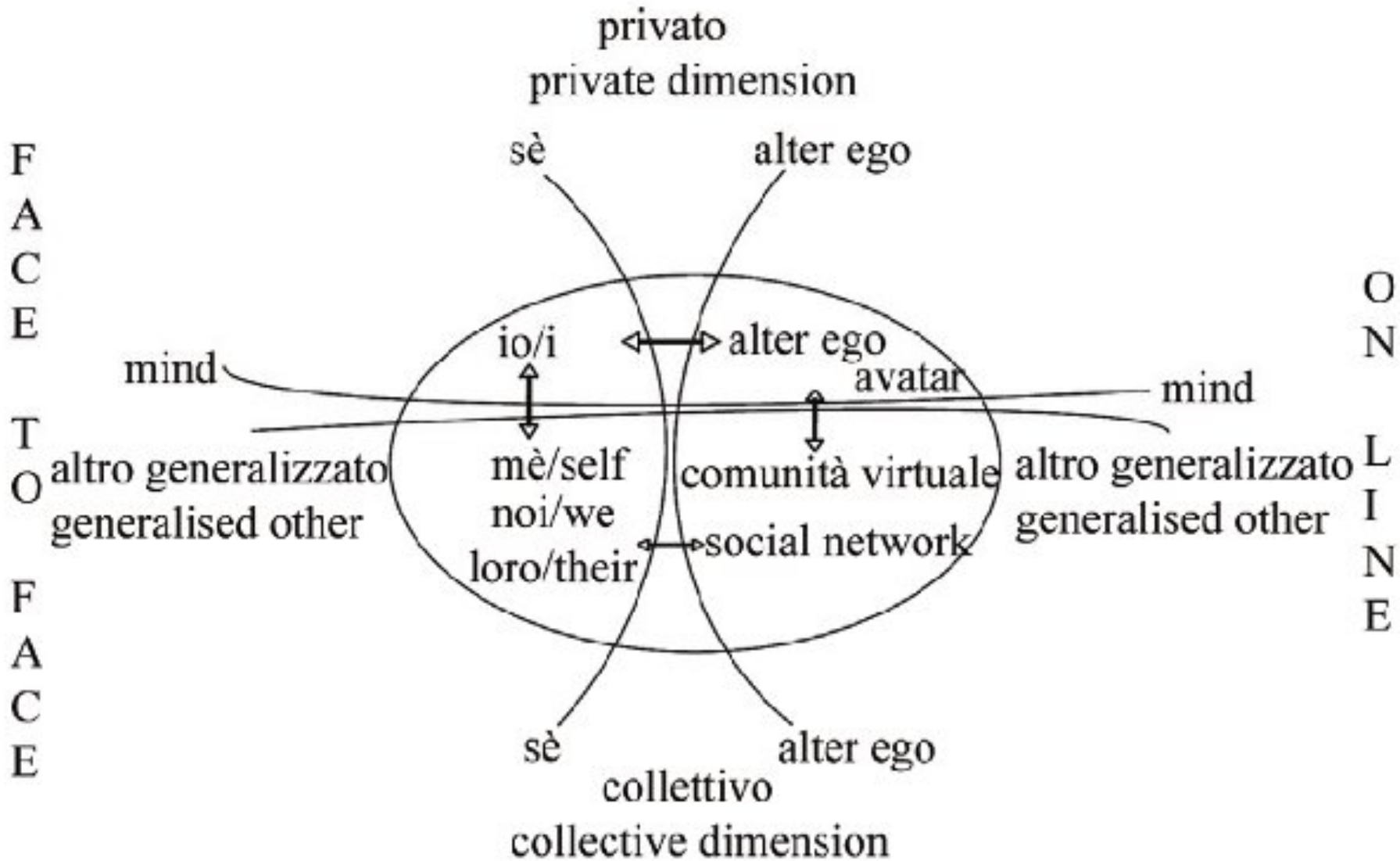
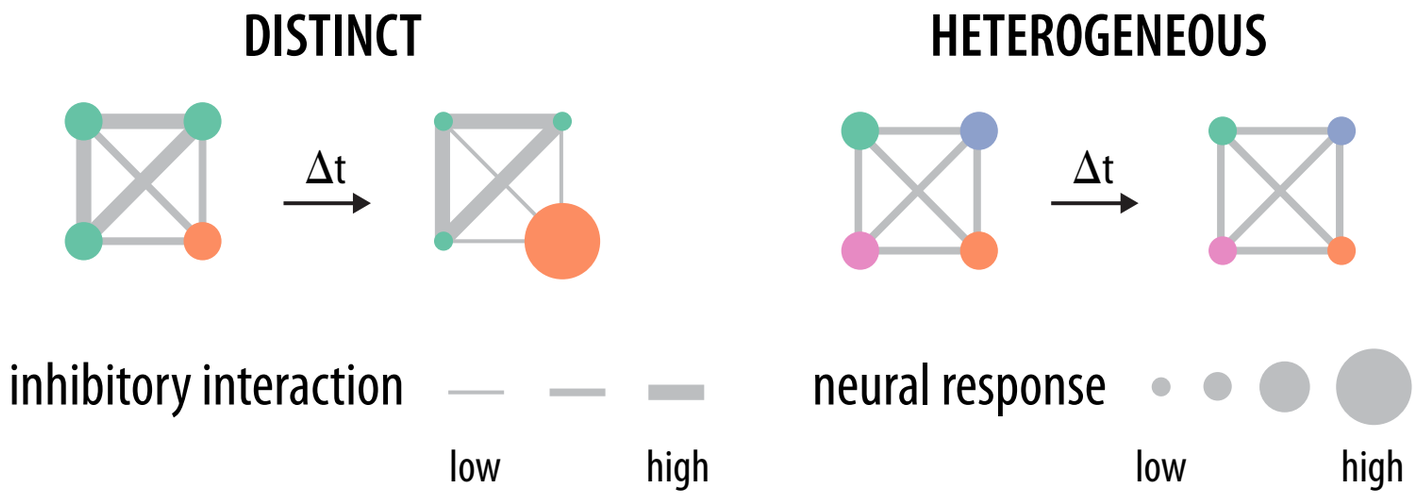


Figure 2.
Severino's flowchart.

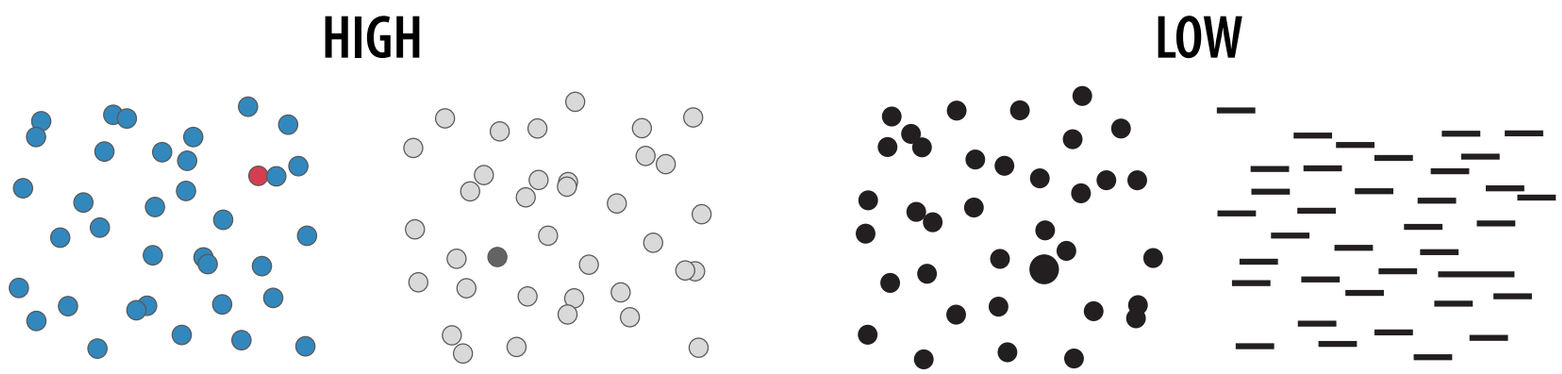
Severino, S., Messina, R. (2011). Group Dynamics in On-Line and Face-to-Face Interactions: An Experimental Study on Learning Methods. Sociology Mind. Vol.1, No.2, 65-73 DOI:10.4236/sm.2011.12008

MATCH VISUAL SALIENCE TO RELEVANCE

ATTENTIONAL CAPTURE



SALIENCE



Fecteau JH, Munoz DP (2006) Salience, relevance, and firing: a priority map for target selection. Trends Cogn Sci 10: 382–390.
Yantis S (2005) How visual salience wins the battle for awareness. Nat Neurosci 8: 975–977.

COLORS COLORS COLORS COLORS COLORS COLORS
COLORS COLORS COLORS COLORS COLORS COLORS
COLORS COLORS COLORS COLORS COLORS COLORS
COLORS COLORS COLORS COLORS COLORS COLORS
COLORS COLORS COLORS COLORS COLORS COLORS
COLORS COLORS COLORS COLORS COLORS COLORS
COLORS COLORS COLORS COLORS COLORS COLORS
COLORS COLORS COLORS COLORS COLORS COLORS
COLORS COLORS COLORS COLORS COLORS COLORS
COLORS COLORS COLORS COLORS COLORS COLORS
COLORS COLORS COLORS COLORS COLORS COLORS
COLORS COLORS COLORS COLORS COLORS COLORS
COLORS COLORS COLORS COLORS COLORS COLORS
COLORS COLORS COLORS COLORS COLORS COLORS

Colours

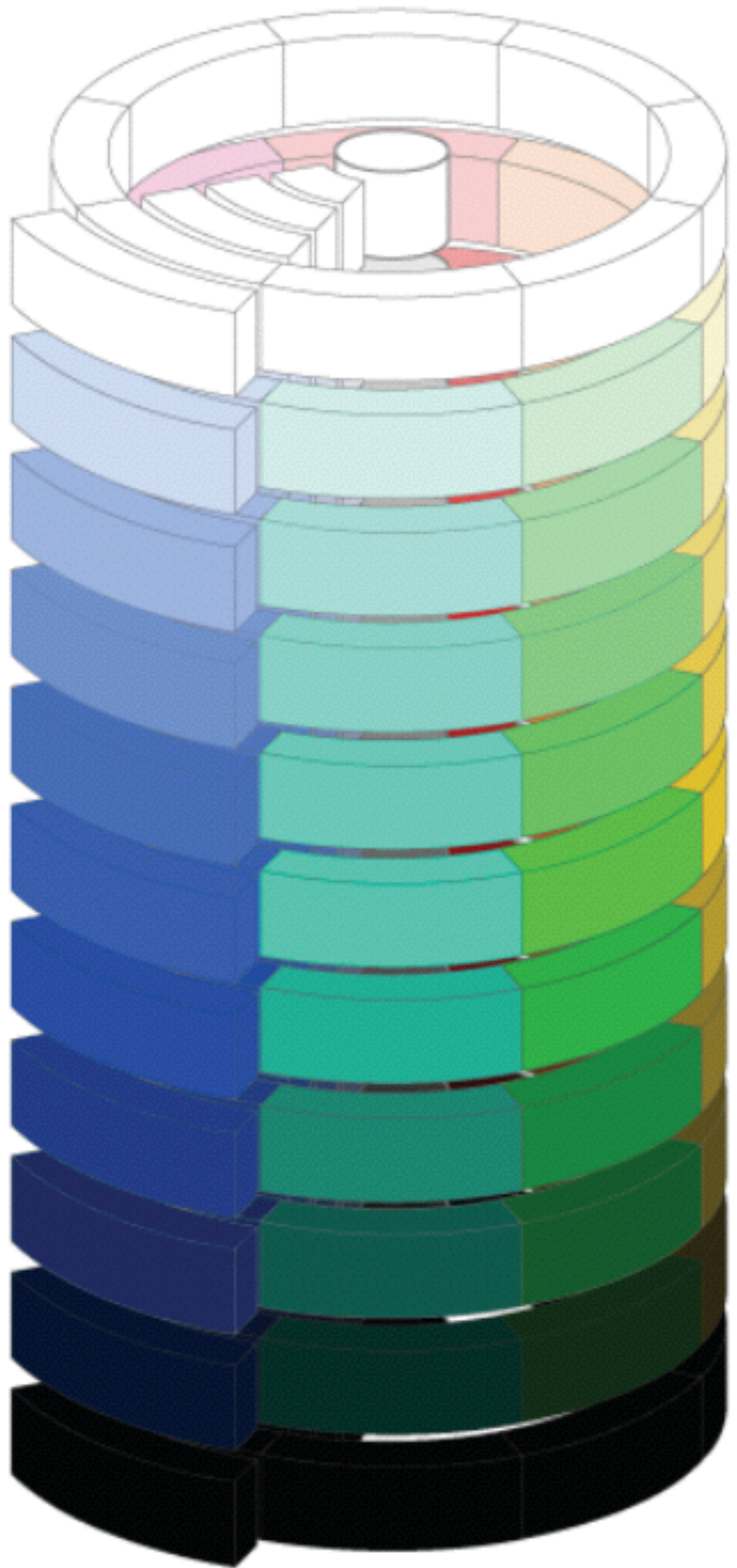
- "Colour used poorly is worse than no colour at all" - Edward Tufte
  - "Above all, do no harm"
  - colour can cause the wrong information to stand out and make meaningful information difficult to see.

Colour space

- A colour space is mathematical model for describing colour.
  - RGB, HSB, HSL, Lab and LCH
- RGB is the most common in computer use,
  - but least useful for design
  - our eyes do not decompose colours into RGBconstituents
- HSV, describes a colour in terms of its hue, saturation and value (lightness),
  - models colour based on intuitive parameters
  - more useful.

Colourimetry

- Hue (colour)
  - around the circle
- Saturation
  - Inside to outside
  - Colour to grey scale
- Lightness (value)
  - top to bottom



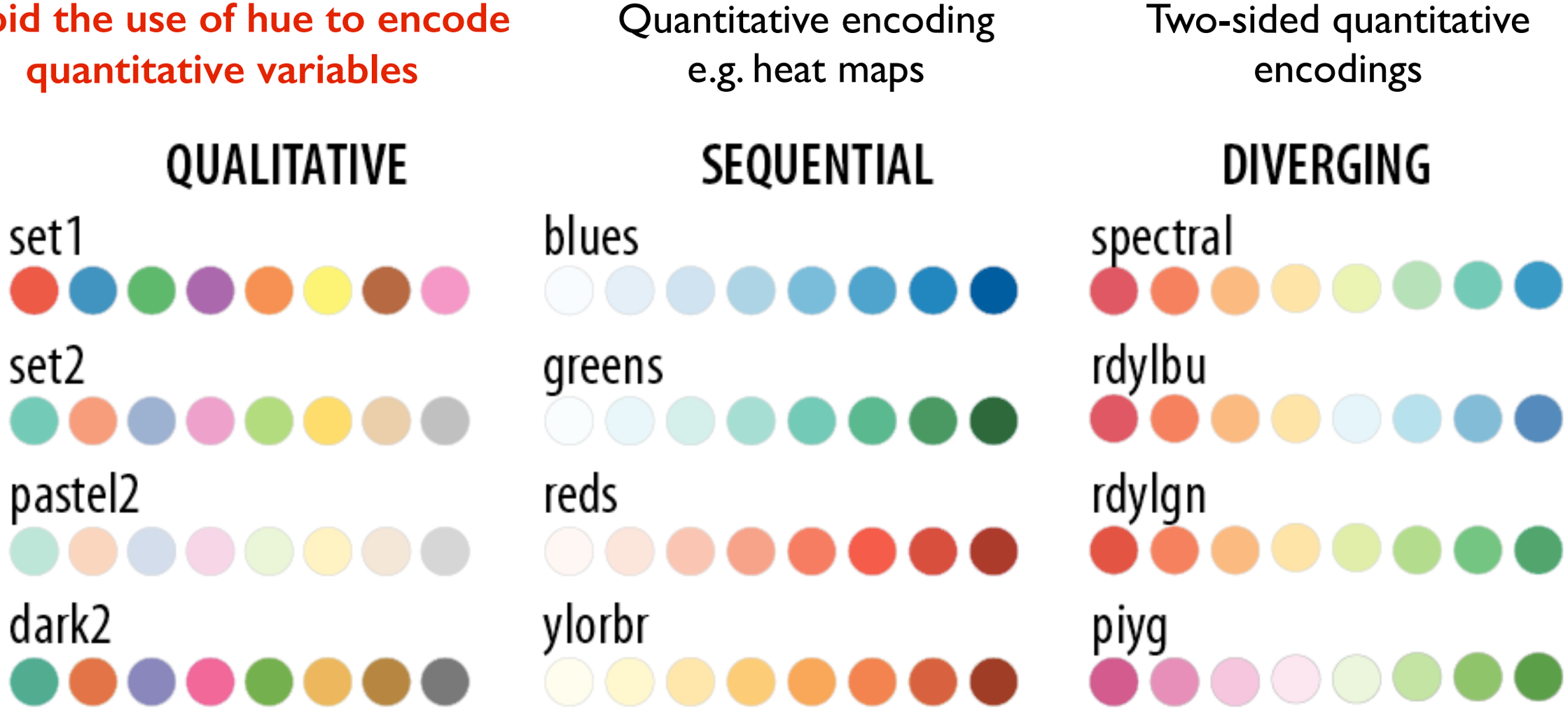
Brewer palettes

- Brewer palettes (colorbrewer.org) provide a range of palettes based on HSV model which make life easier for us….

Avoid the use of hue to encode quantitative variables

Quantitative encoding e.g. heat maps

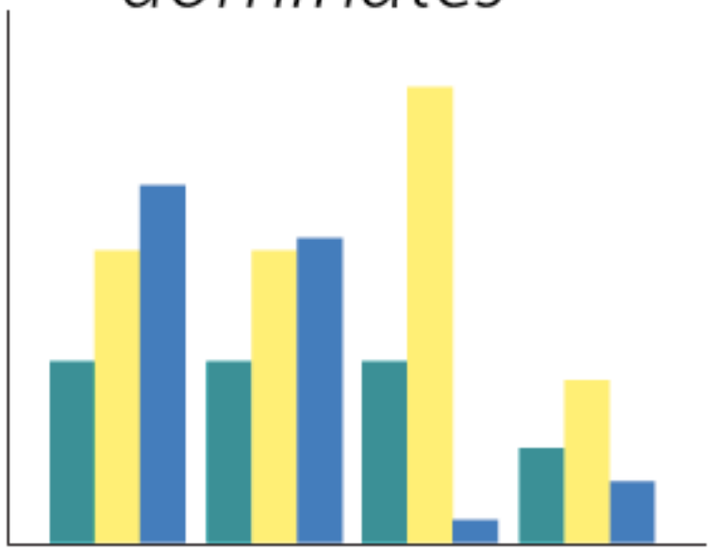
Two-sided quantitative encodings



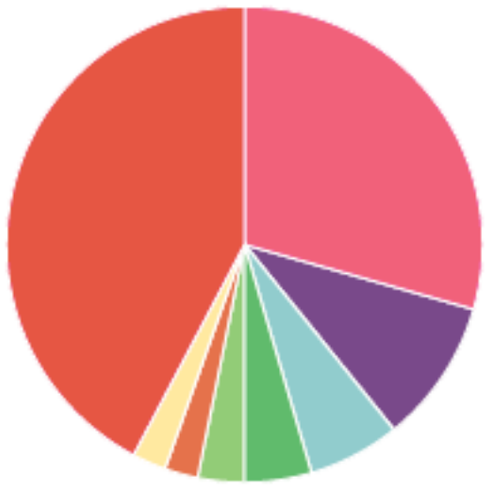
Examples

Poor use of colour

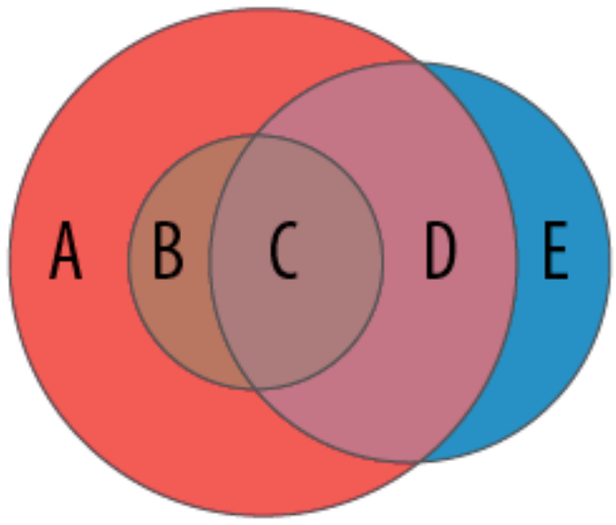
one color dominates



difficult to distinguish

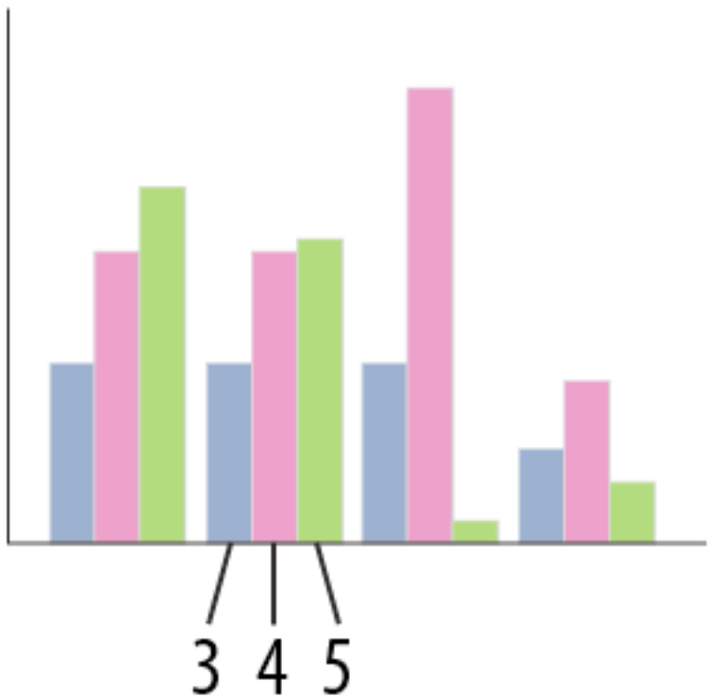


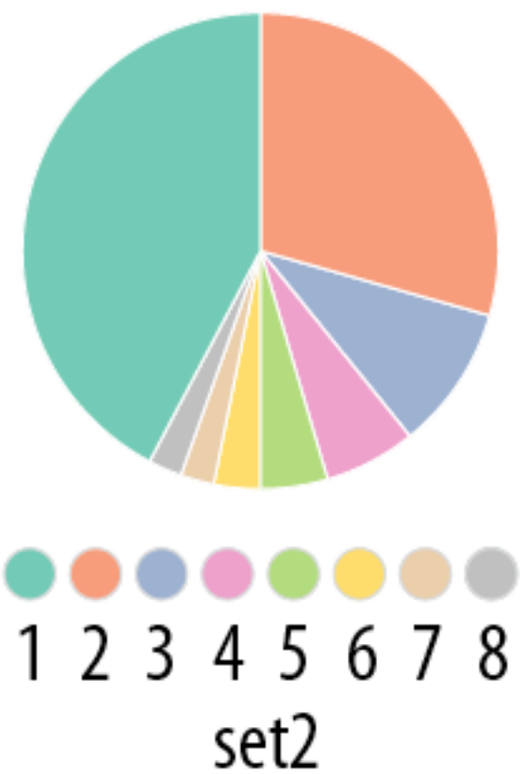
murky

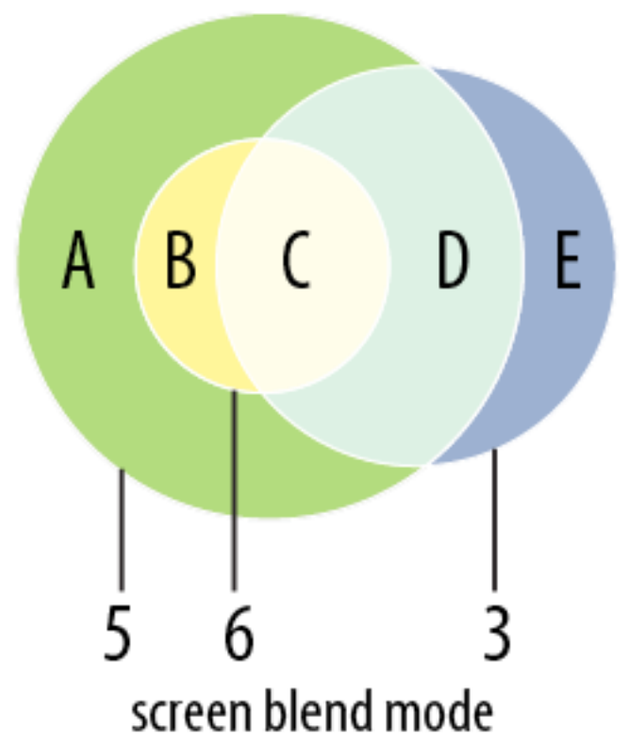


Examples

recolored with Brewer palettes

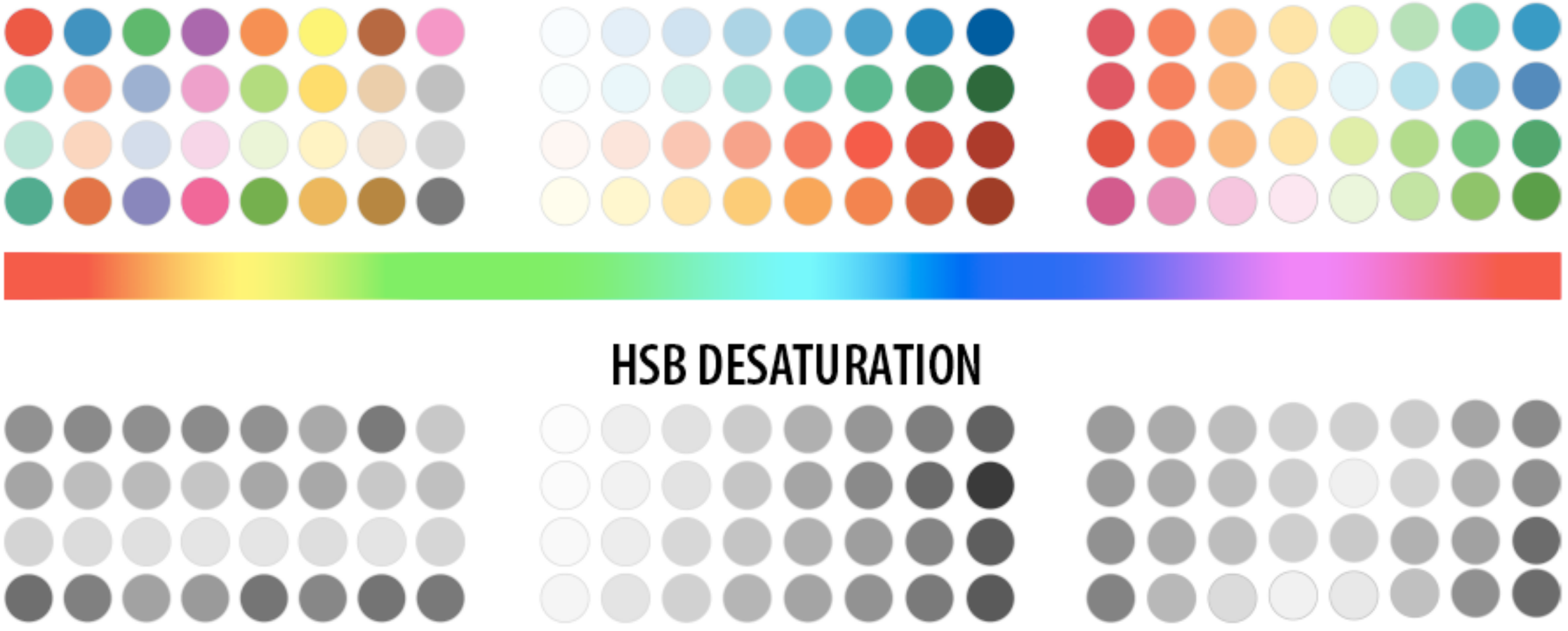




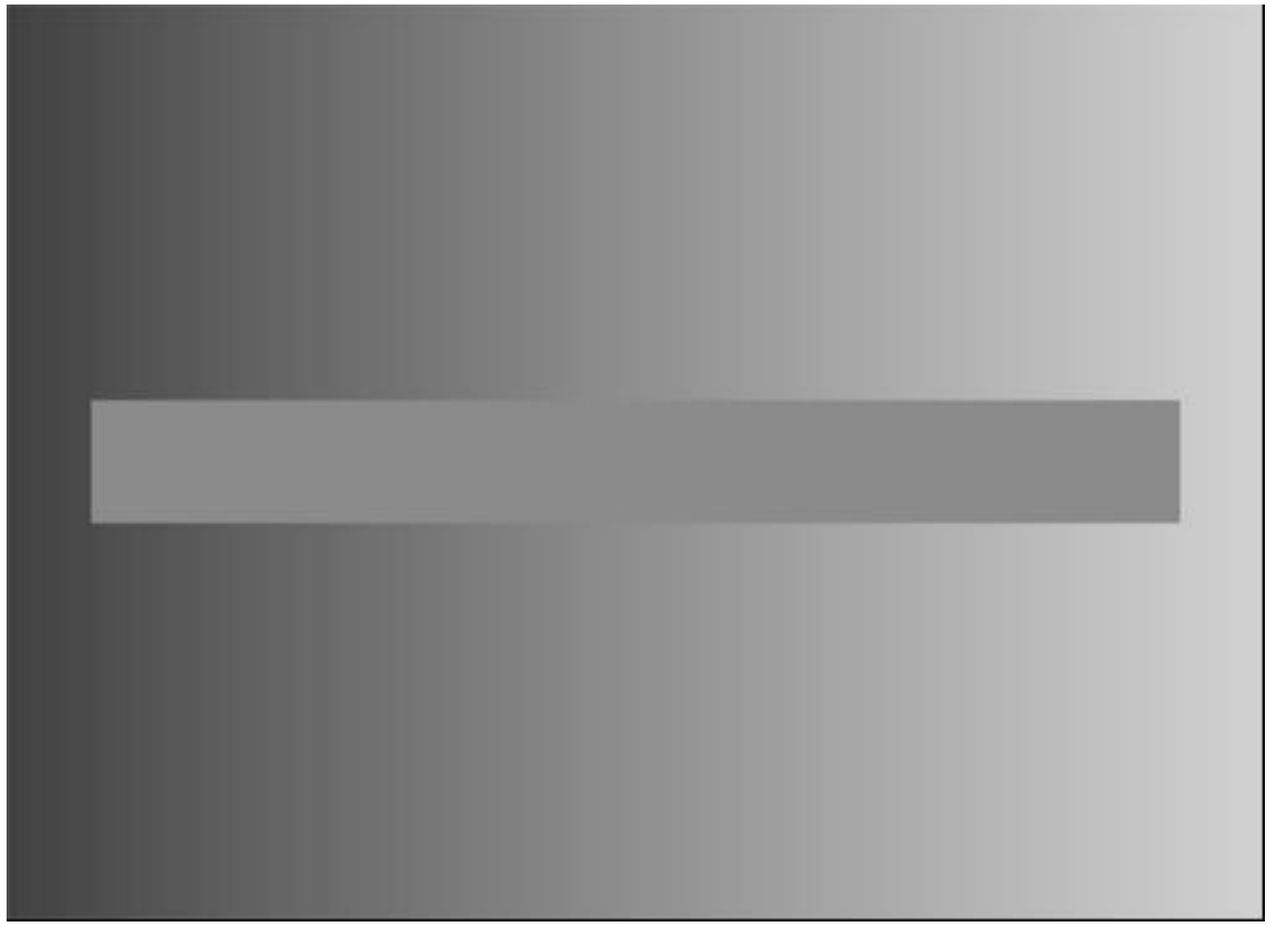


Conversion to Grey scale

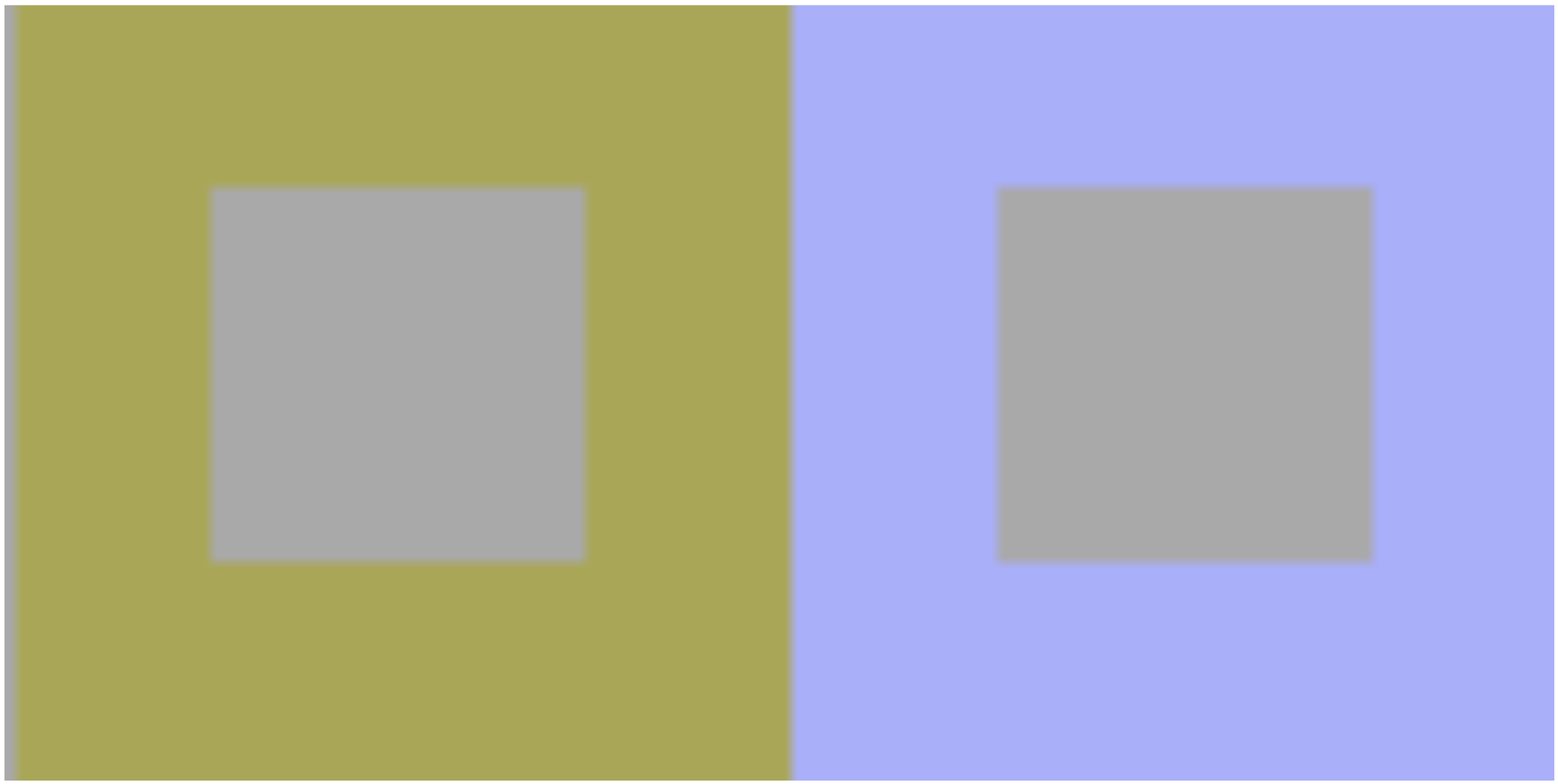
- Ensure chosen colour set works well in grey scale
  - Sequential palette works well here



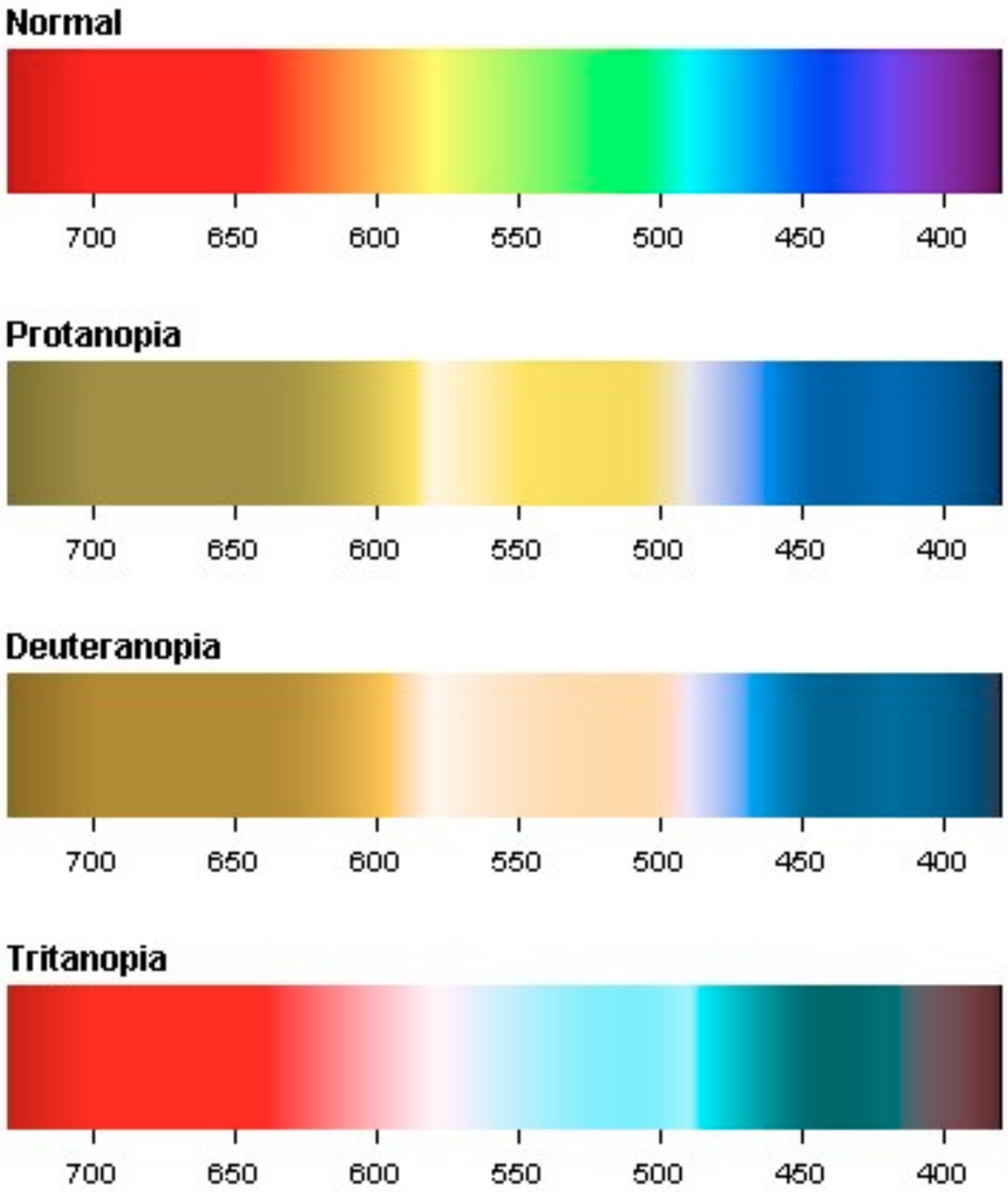
Trouble with perceptual colour….

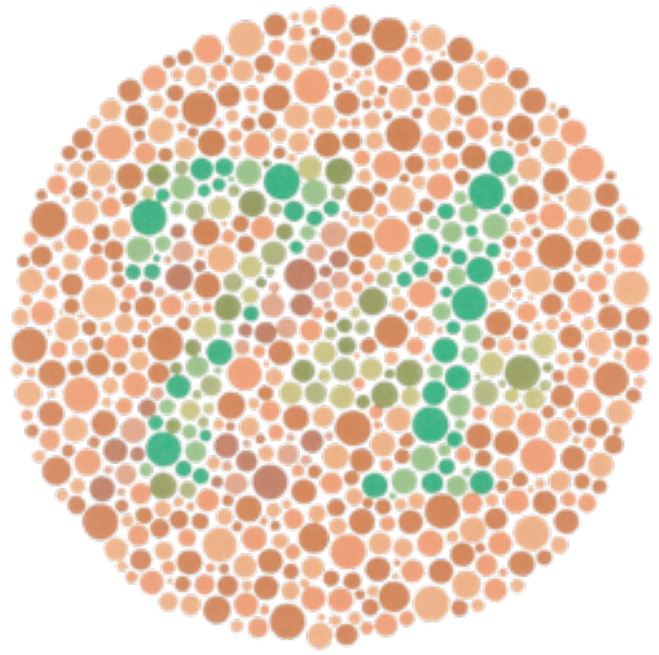


Context Affects Perceived Colour



Colour & Accessibility





Accessibility (W3C): 10-20% of population are red/green colour blind.

Color blindness

red-green

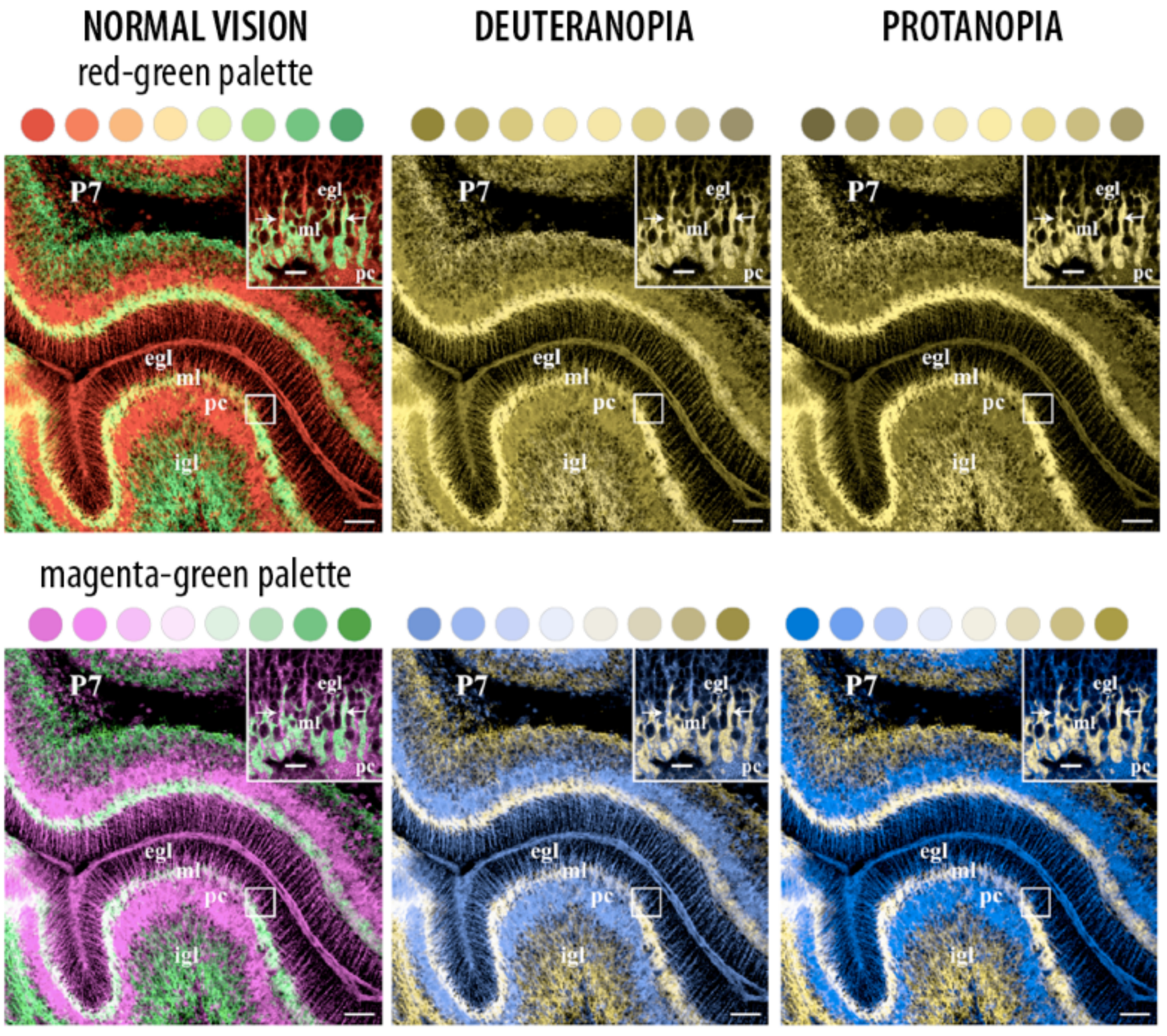
red-green

blue-yellow

BioVis Example: Immunofluorescence images

red-green image of P2Y1 receptor and migrating granule neurons,

effectively remapped to magenta-green using the channel mixing method.

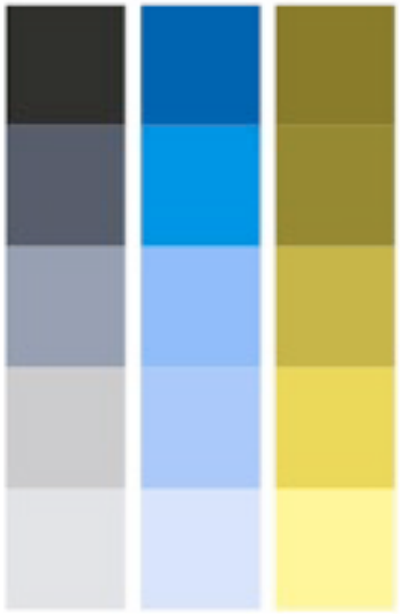


15-COLOR PALETTE FOR COLOR BLINDNESS

NORMAL VISION

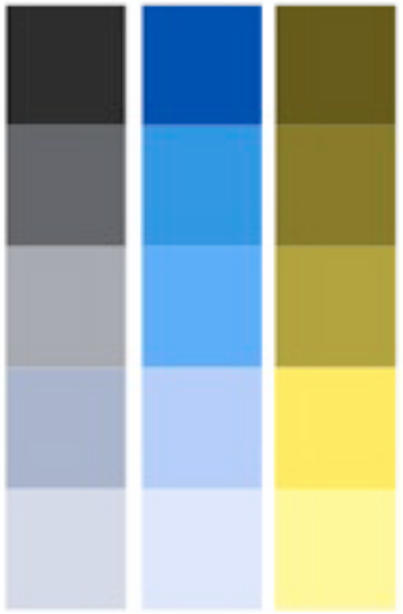
DEUTERANOPIA

common (6%)



PROTANOPIA

rare (2%)

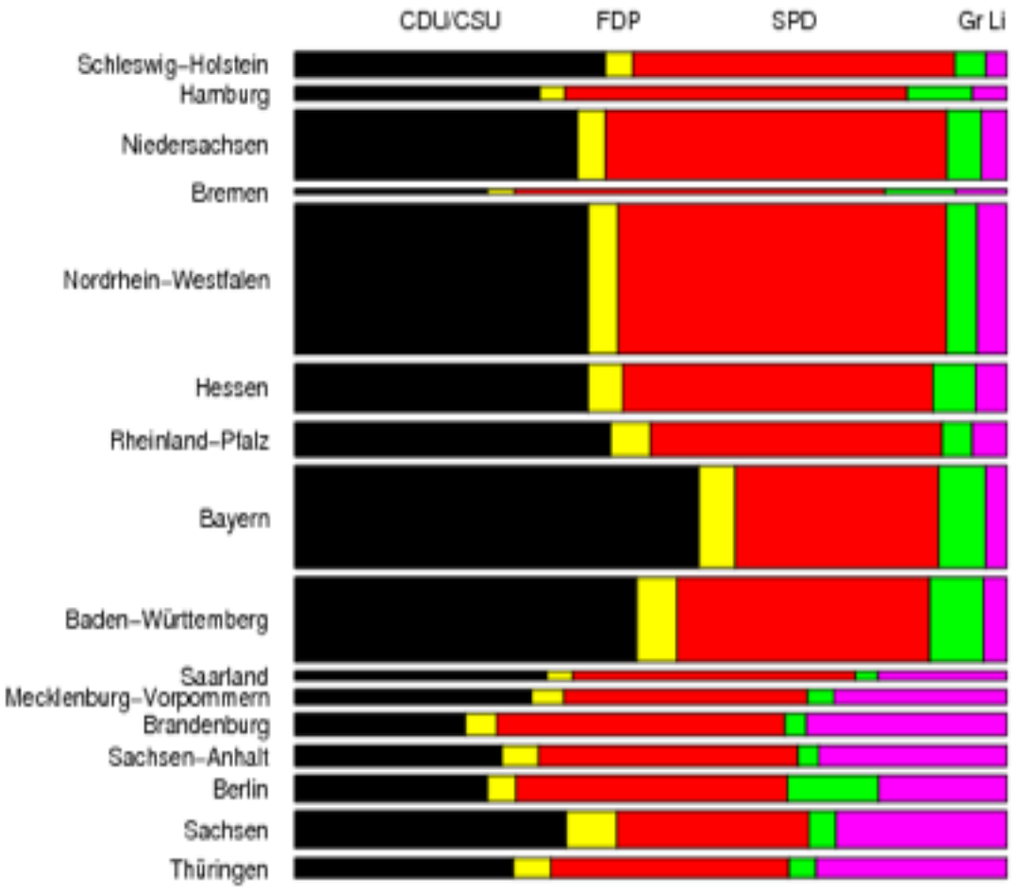


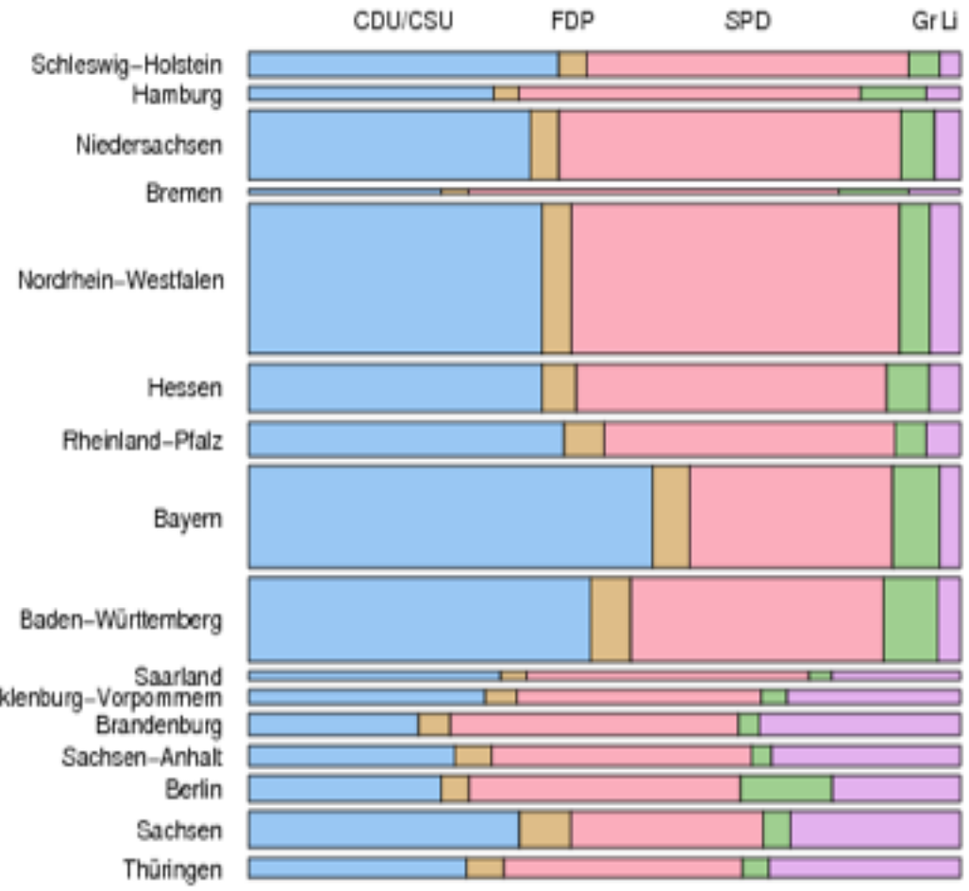
TRITANOPIA

very rare (<1%)

	R	G	B		R	G	B		R	G	B
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

Biased and unbiased politics

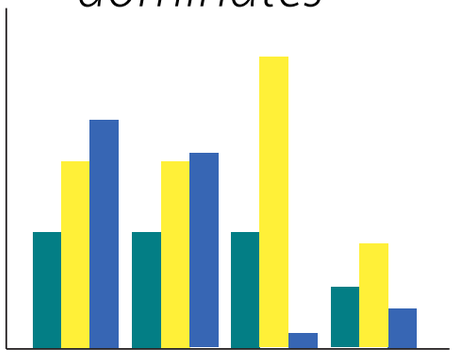




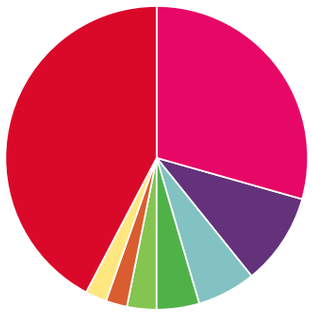
From A. Zeileis, Reisensburg 2007

USE BREWER PALETTES

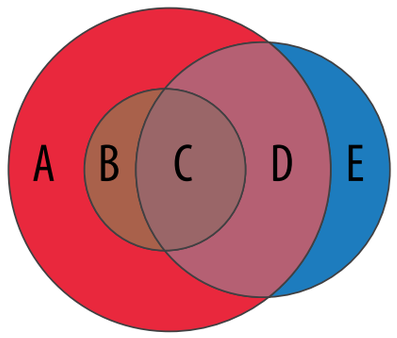
one color dominates



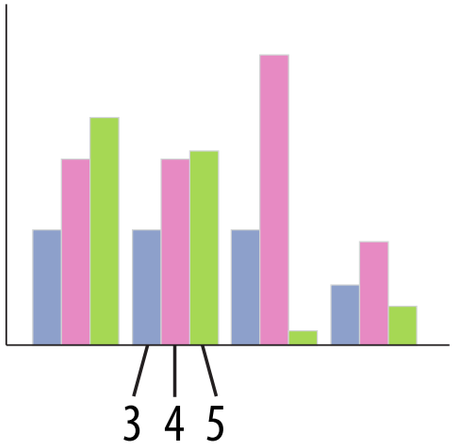
difficult to distinguish

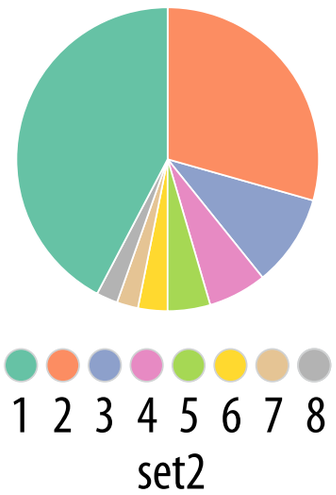


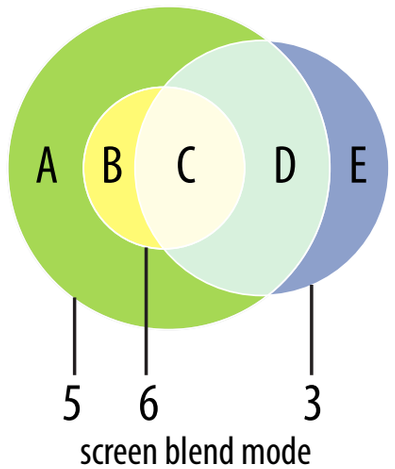
murky



recolored with Brewer palettes



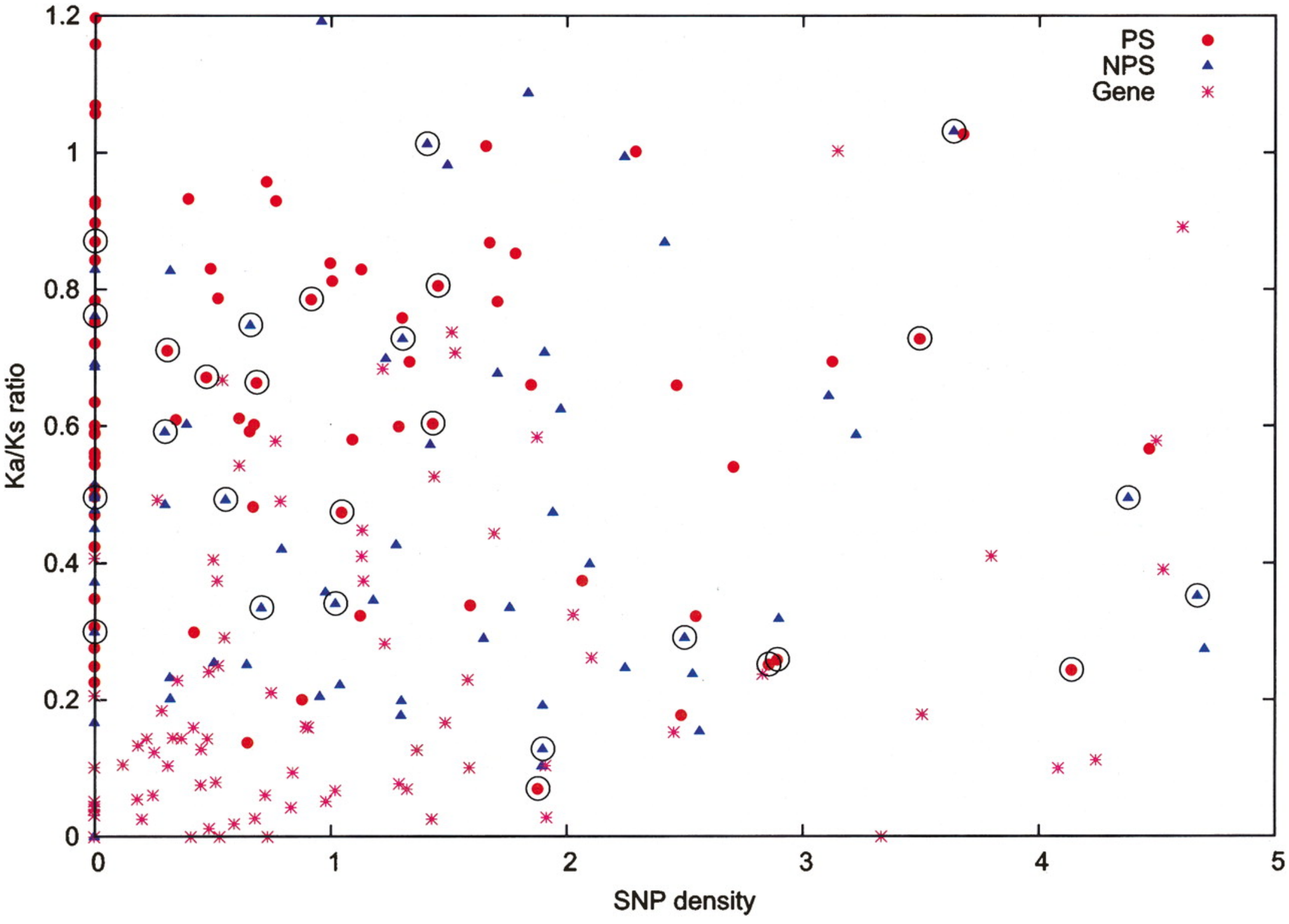




Vettore AL, da Silva FR, Kemper EL, Souza GM, da Silva AM, et al. (2003) Analysis and functional annotation of an expressed sequence tag collection for tropical crop sugarcane. Genome Res 13: 2725–2735.
Bono H, Yagi K, Kasukawa T, Nikaido I, Tominaga N, et al. (2003) Systematic expression profiling of the mouse transcriptome using RIKEN cDNA microarrays. Genome Res 13: 1318–1323.
Tenney AE, Wu JQ, Langton L, Klueh P, Quatrano R, et al. (2007) A tale of two templates: automatically resolving double traces has many applications, including efficient PCR–based elucidation of alternative splices. Genome Res 17: 212–218.

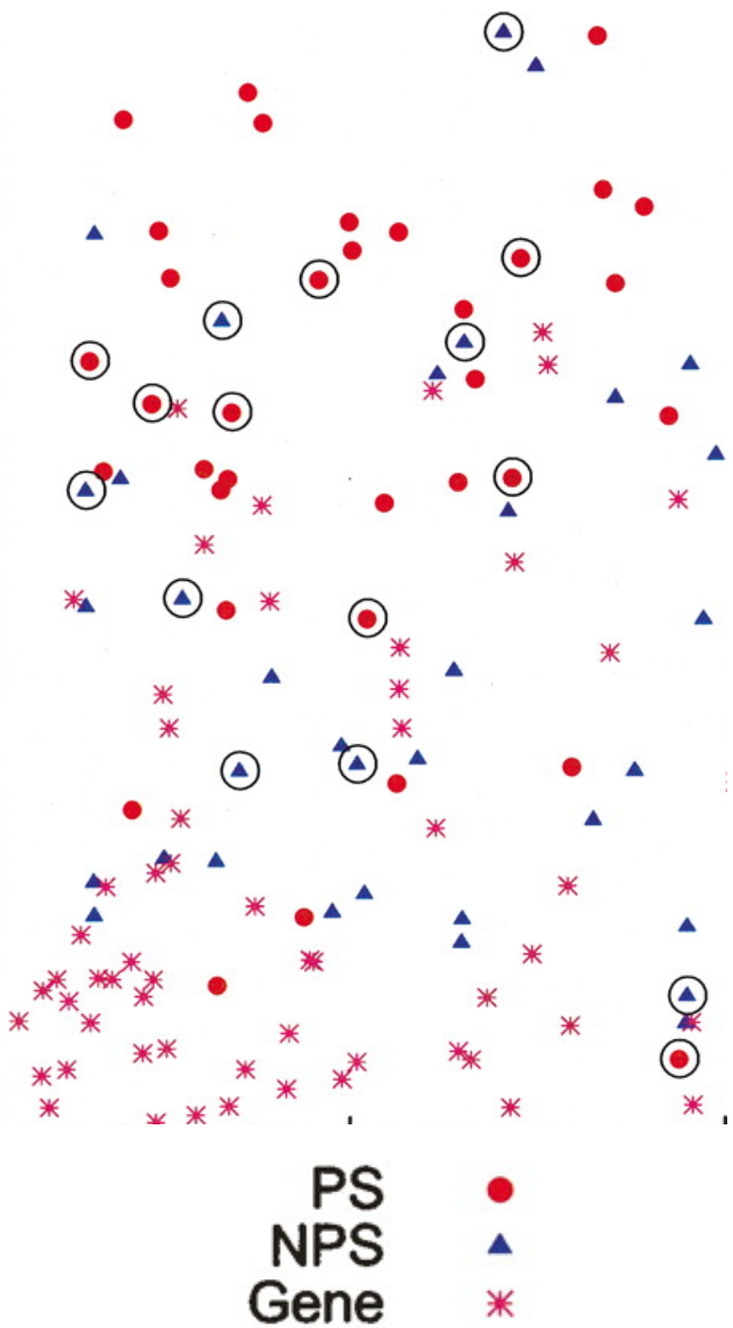
HIERARCHY AND PRIORITY

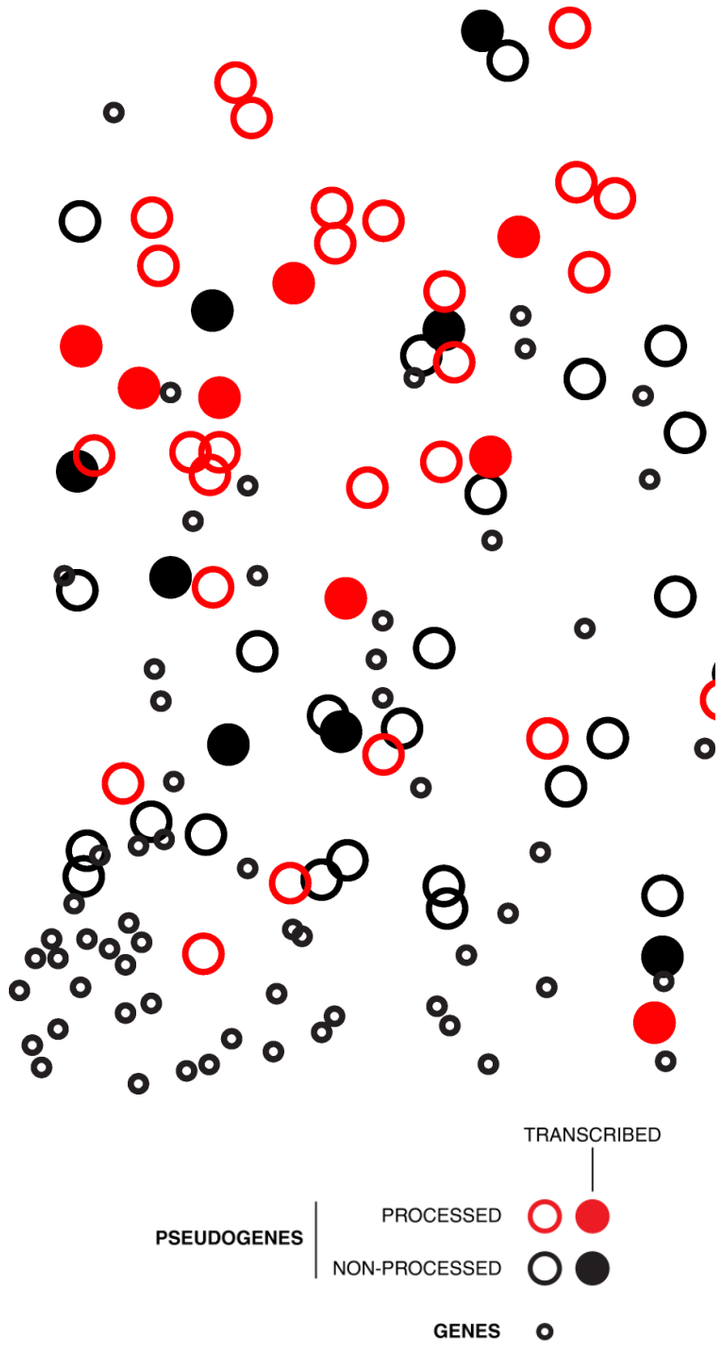
Use symbols that intuitively encode related concepts.



Zheng, D., et al., Pseudogenes in the ENCODE regions: consensus annotation, analysis of transcription, and evolution. Genome Res, 2007. 17(6): p. 839-51.

Zheng, D., et al., Pseudogenes in the ENCODE regions: consensus annotation, analysis of transcription, and evolution. Genome Res, 2007. 17(6): p. 839-51.

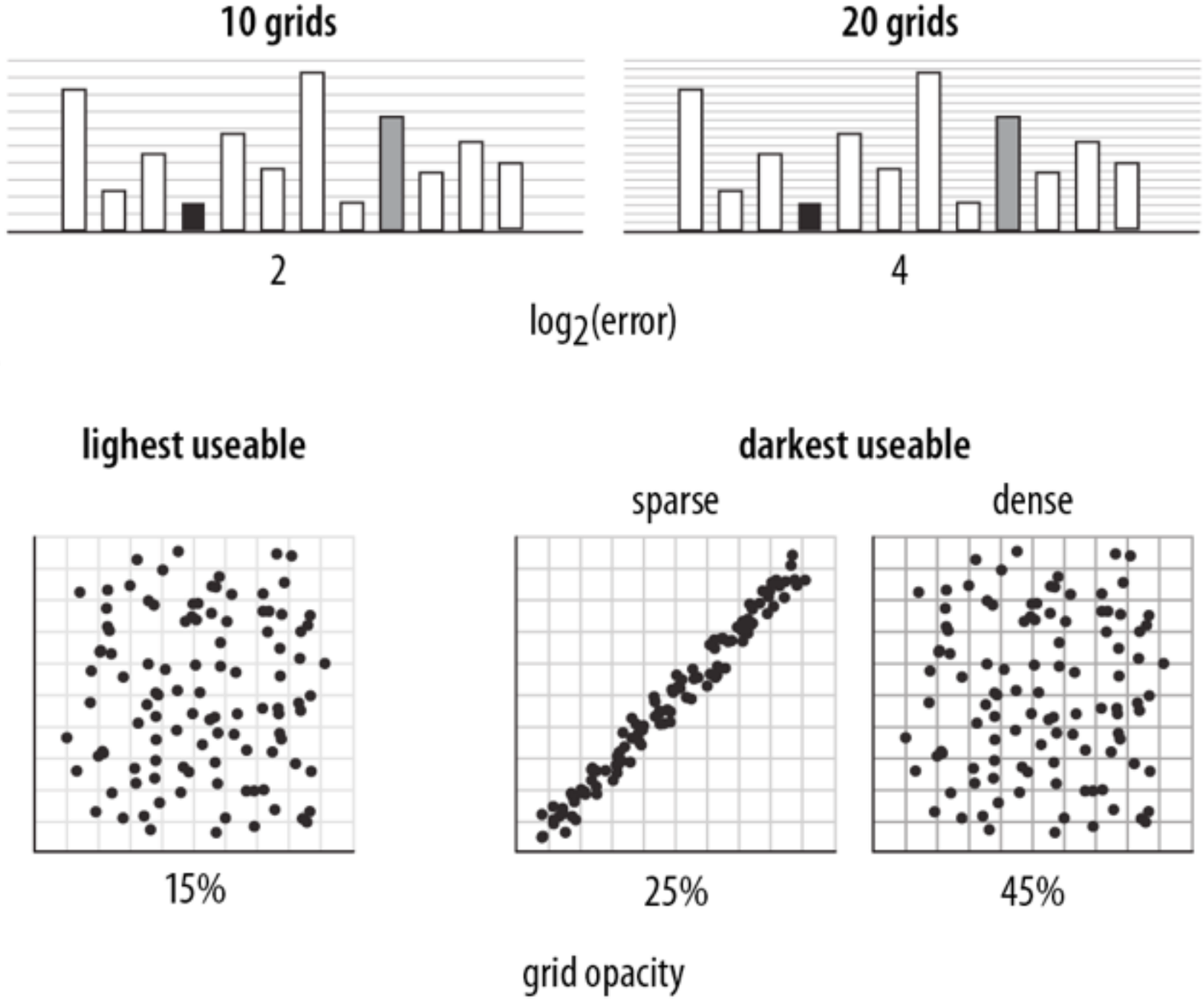




Zheng, D., et al., Pseudogenes in the ENCODE regions: consensus annotation, analysis of transcription, and evolution. Genome Res, 2007. 17(6): p. 839-51.

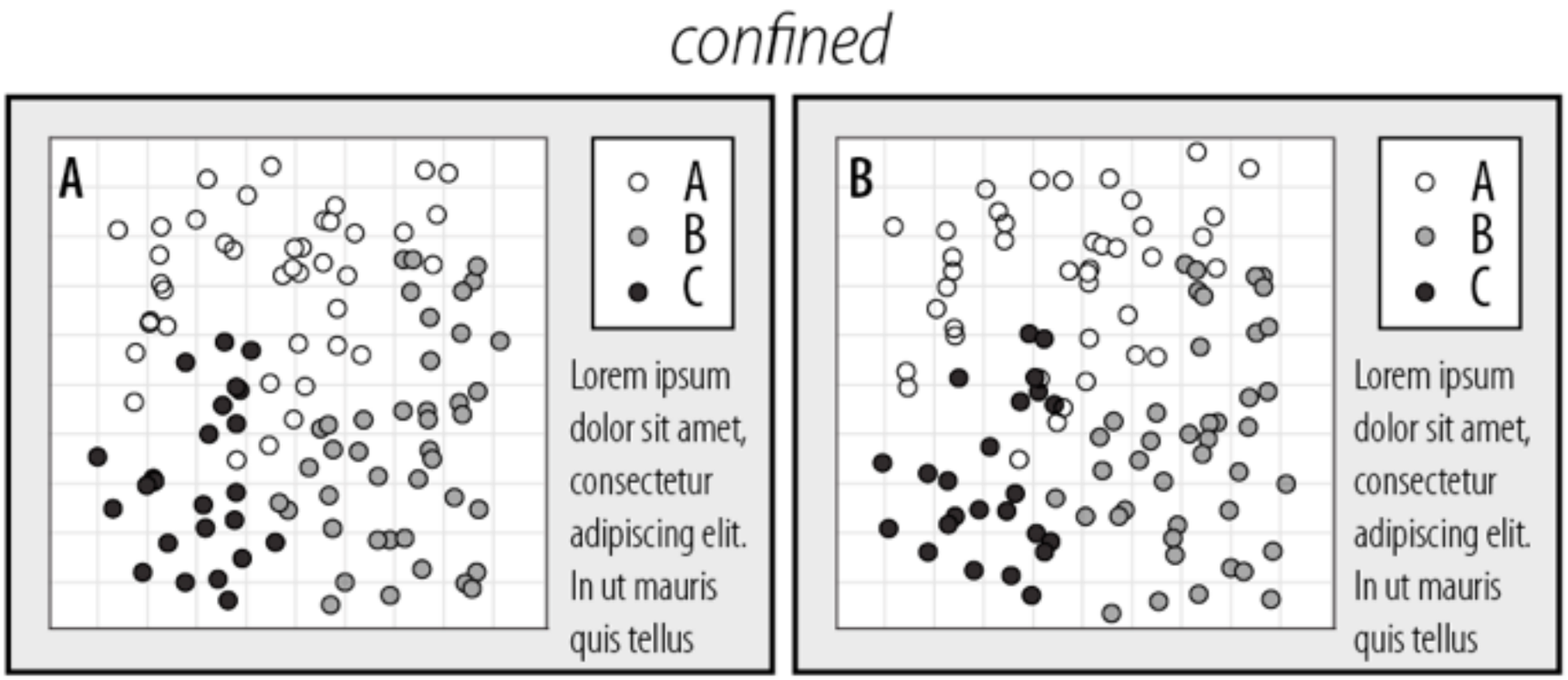
Increase data:ink ratio

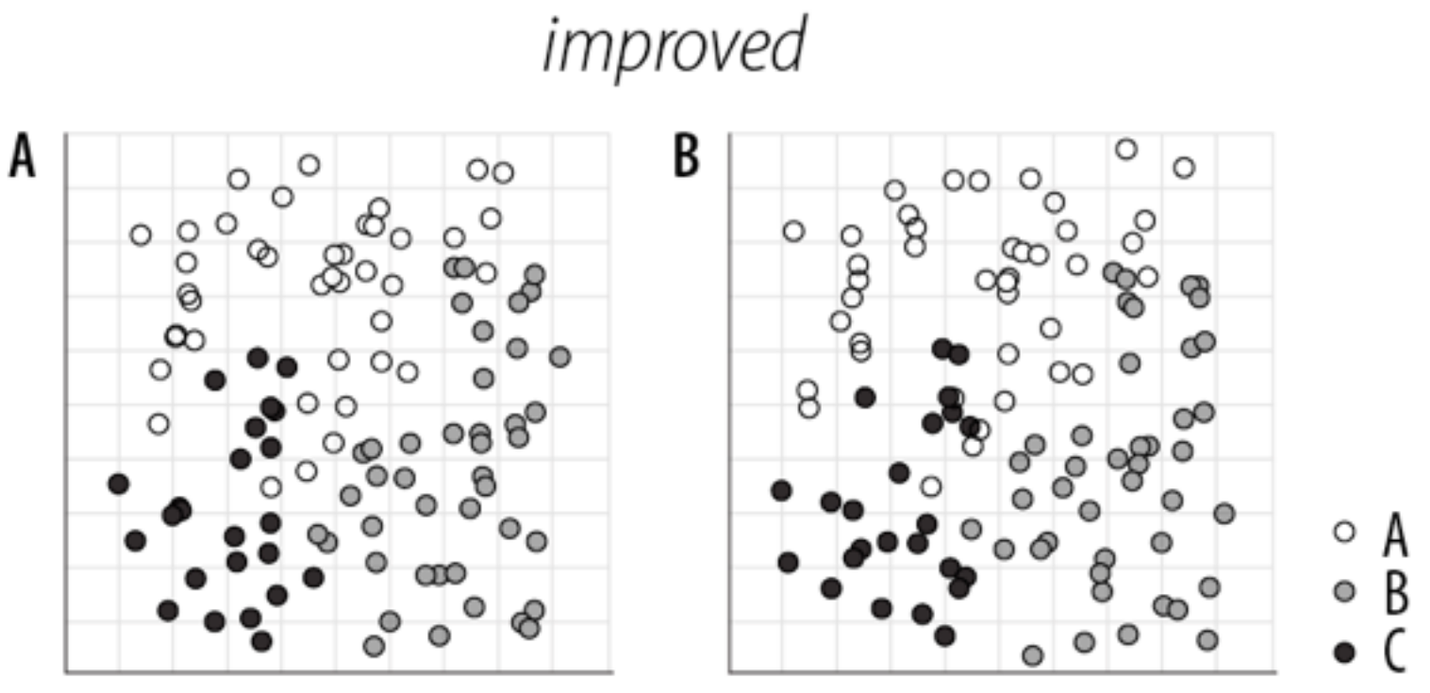
- Navigational aids
  - should not compete with the data for salience.
- Avoid
  - heavy axes
  - error bars and
  - glyphs



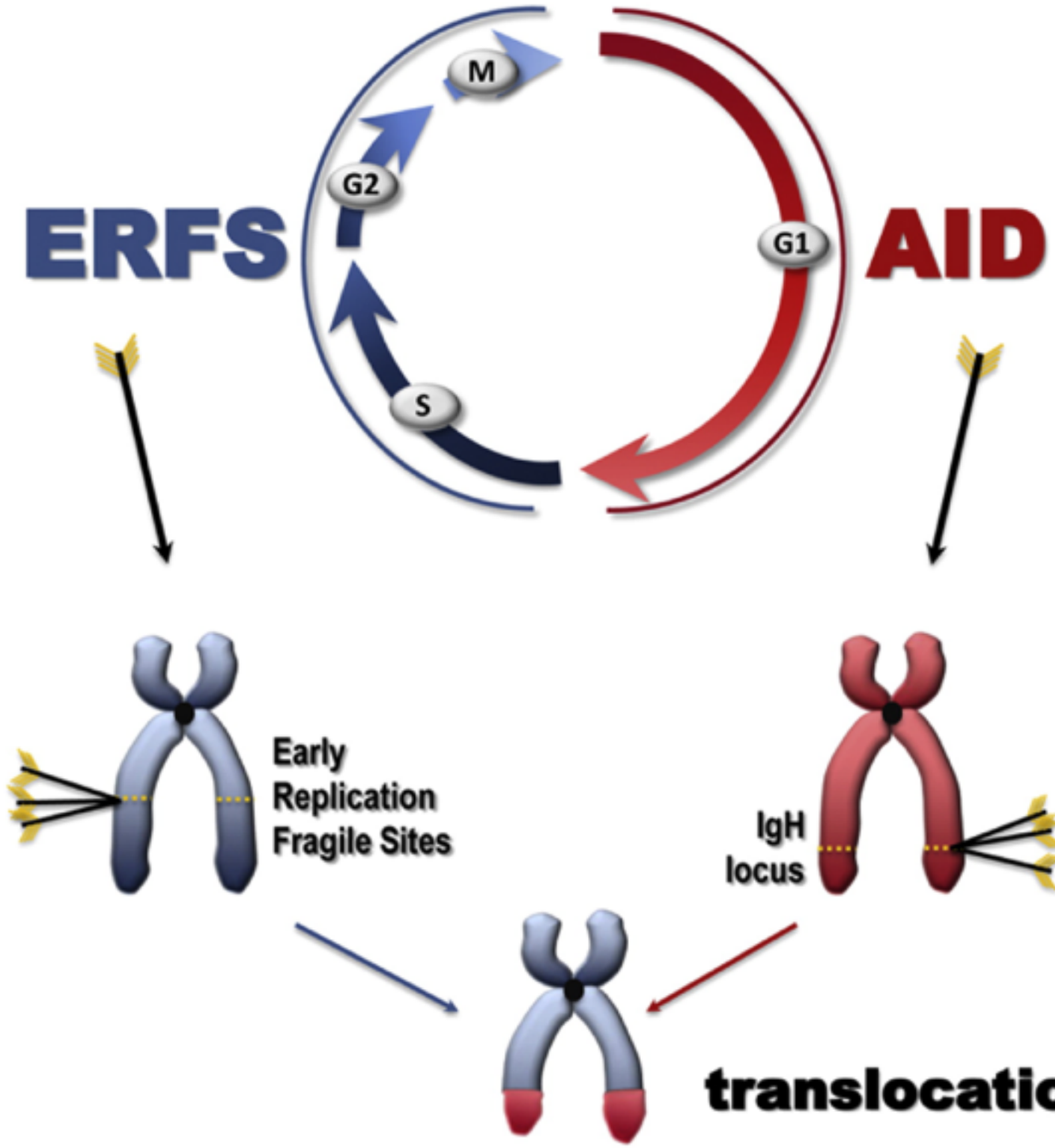
Increase data:ink ratio

- Avoid unnecessary containment

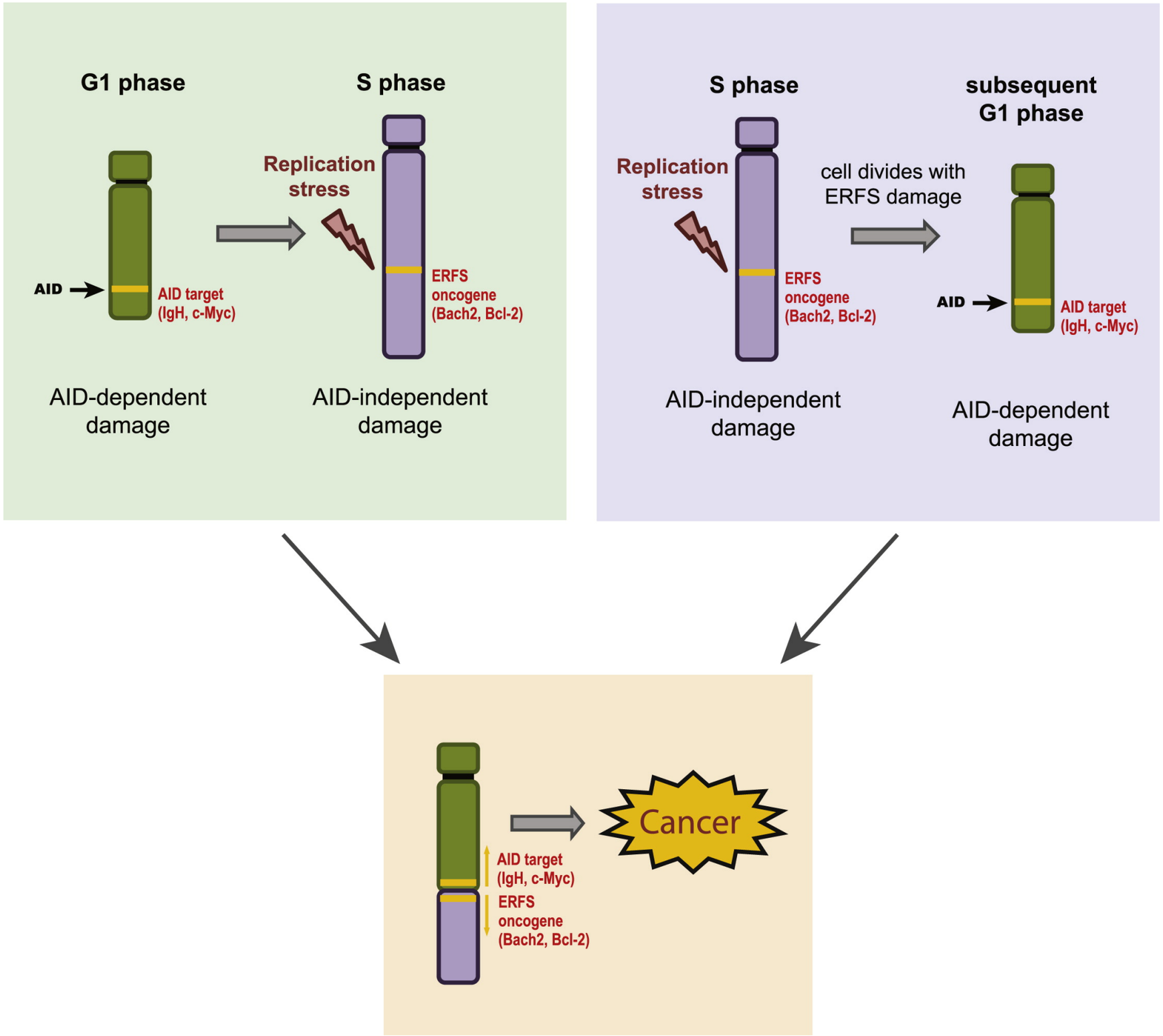




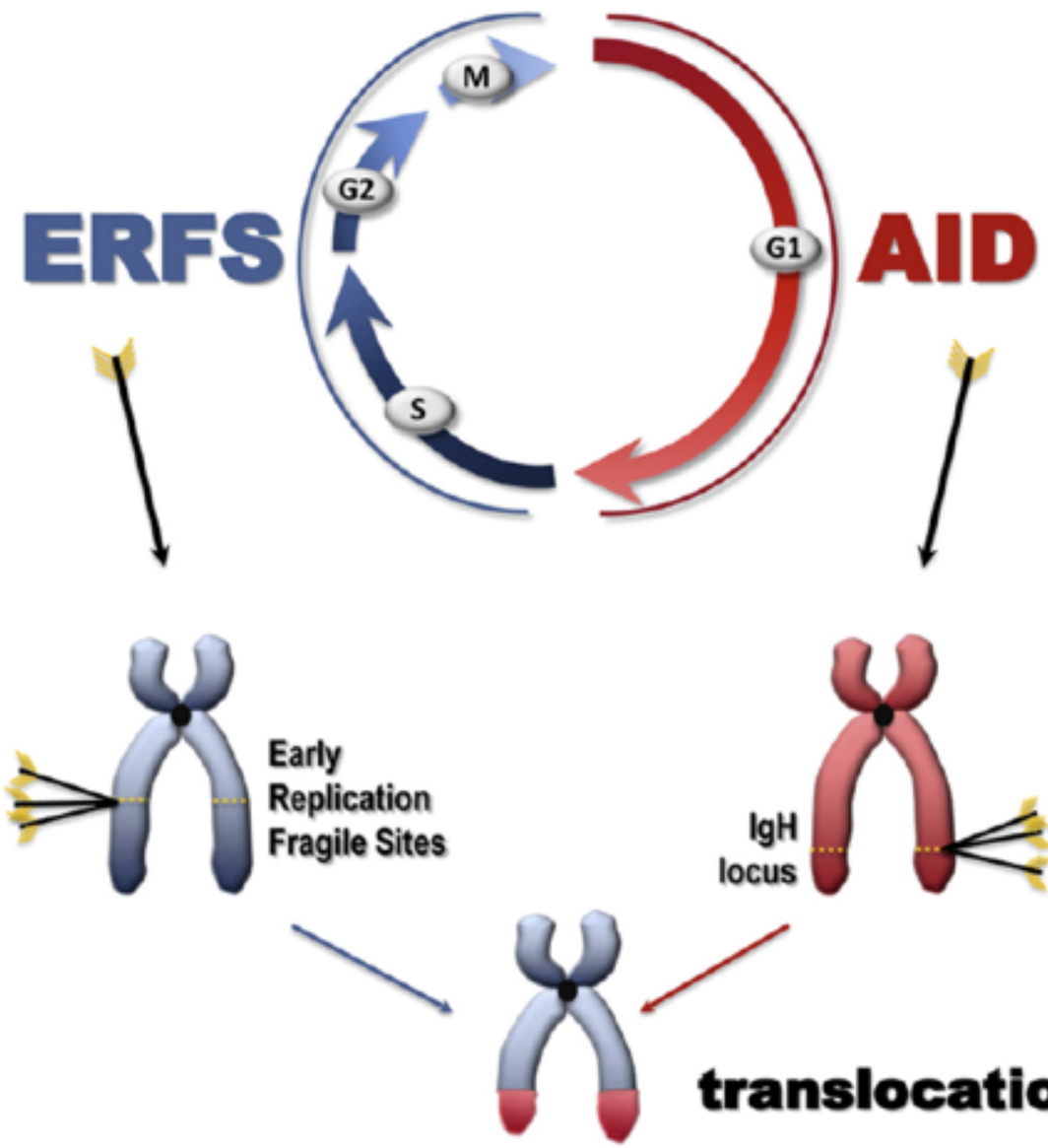
VISUAL ABSTRACT REDESIGN

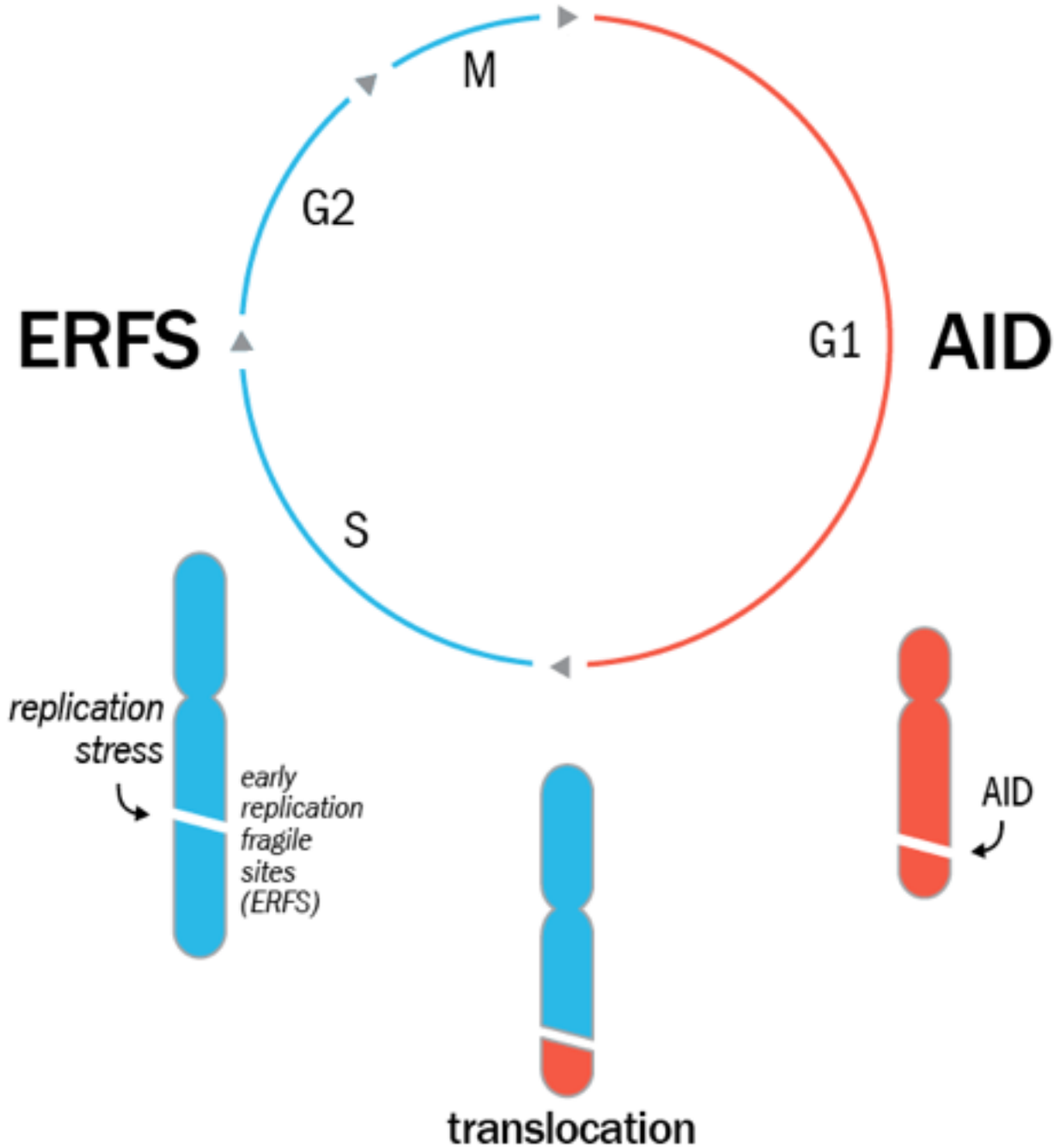


JH Barlow et al (2013) Identification of Early Replicating Fragile Sites that Contribute to Genome Instability. Cell 152:620–632.

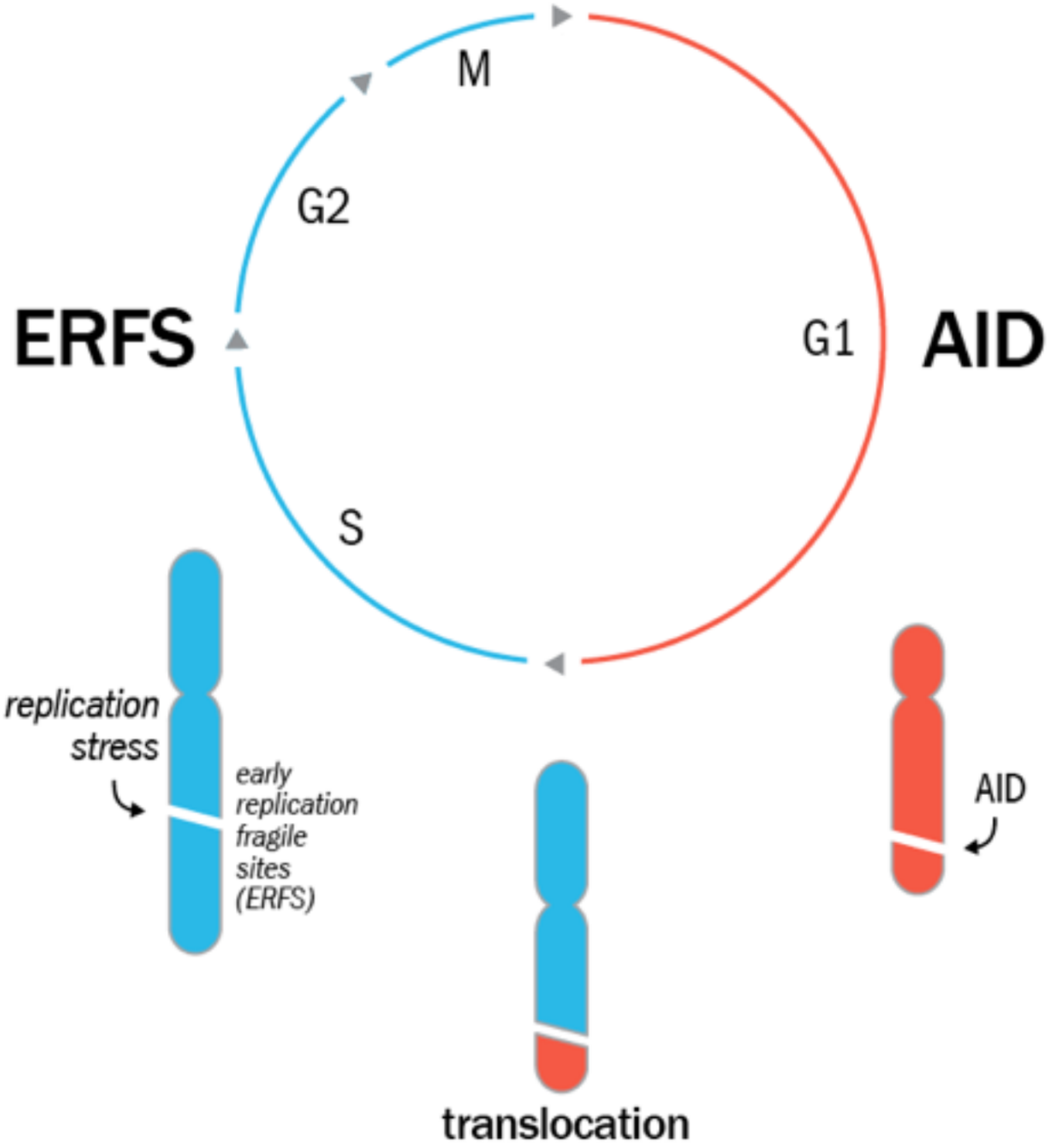


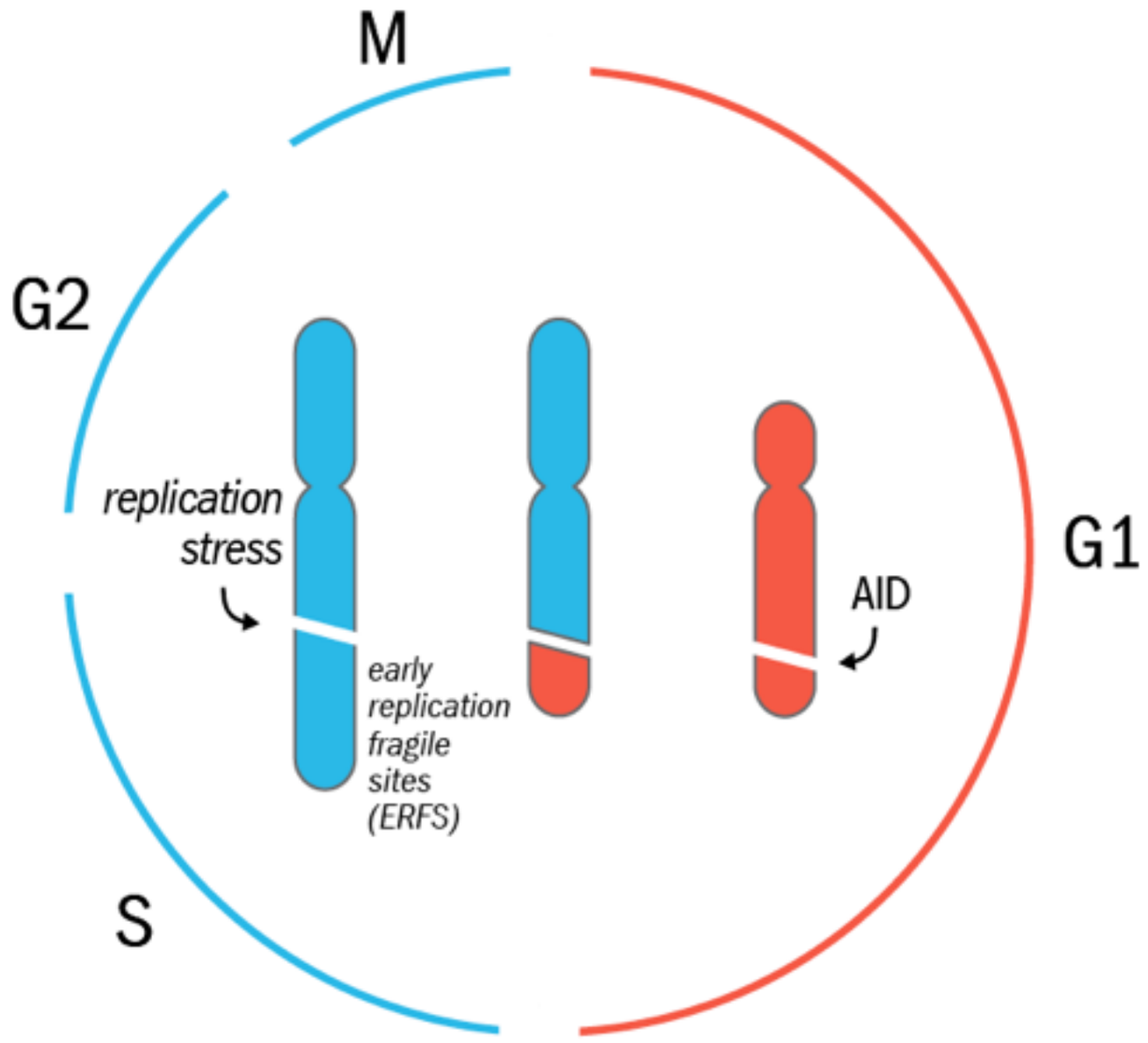
JH Barlow et al (2013) Identification of Early Replicating Fragile Sites that Contribute to Genome Instability. Cell 152:620–632.



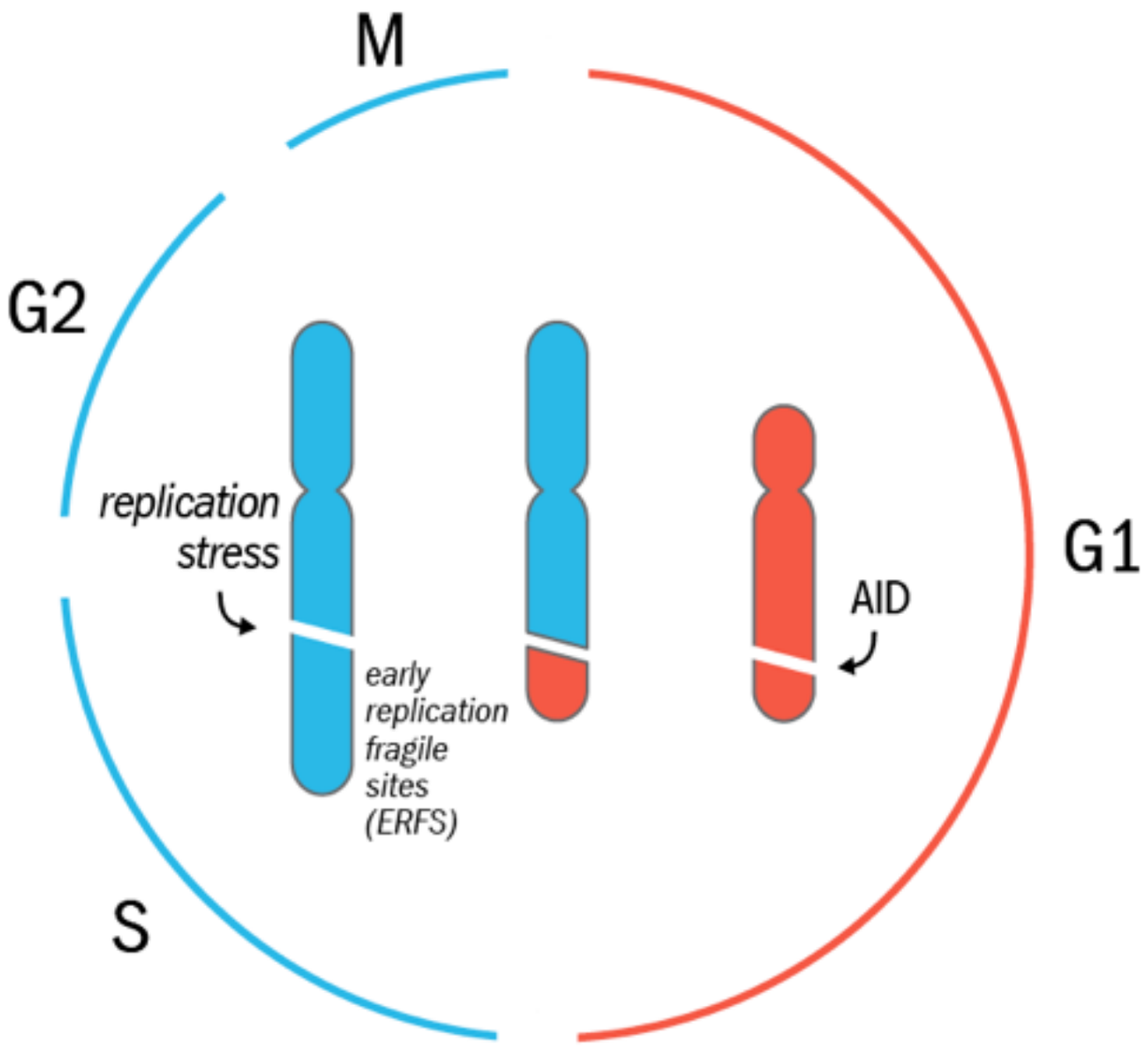


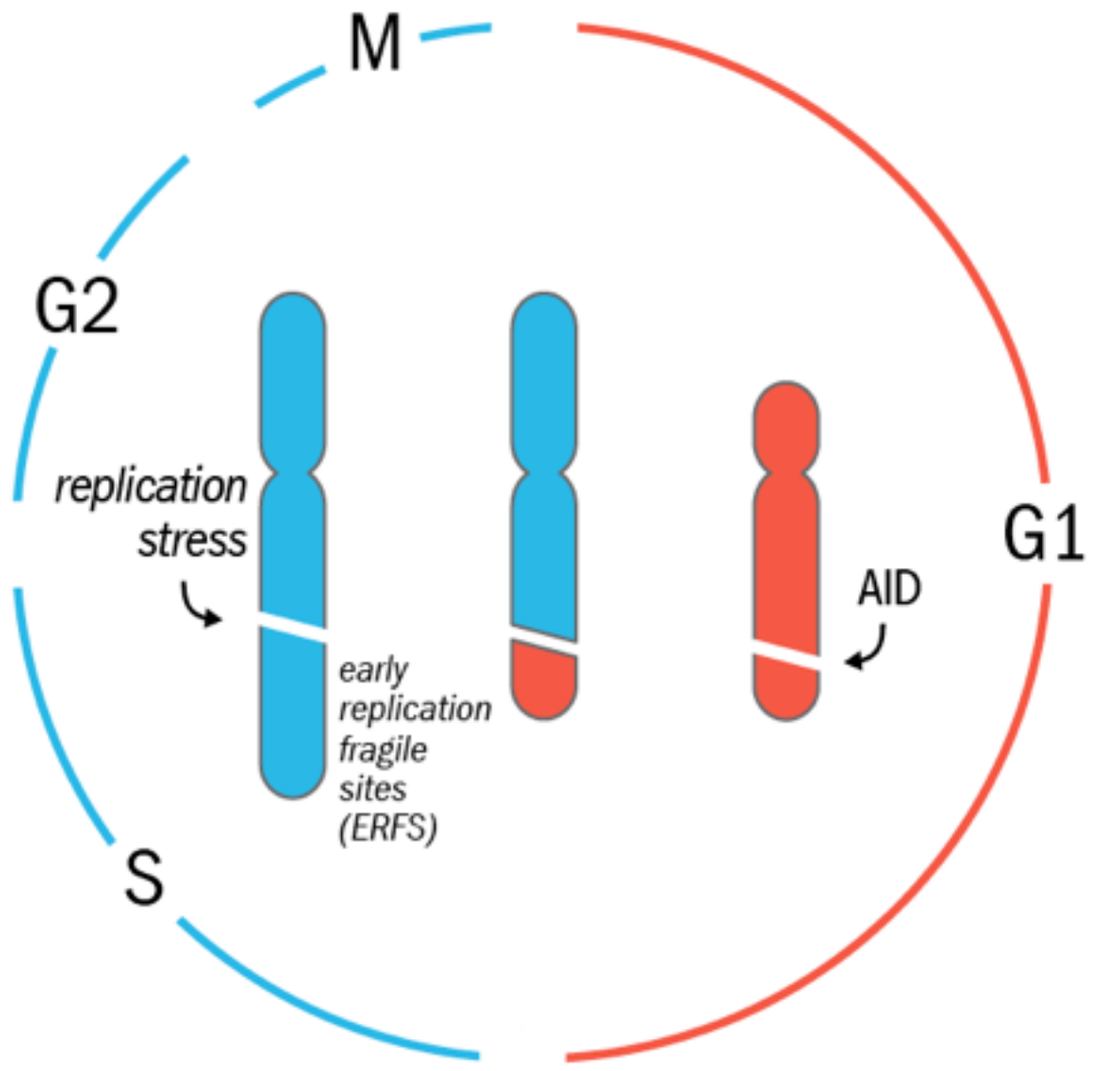
JH Barlow et al (2013) Identification of Early Replicating Fragile Sites that Contribute to Genome Instability. Cell 152:620–632.
redesign by M Krzywinski



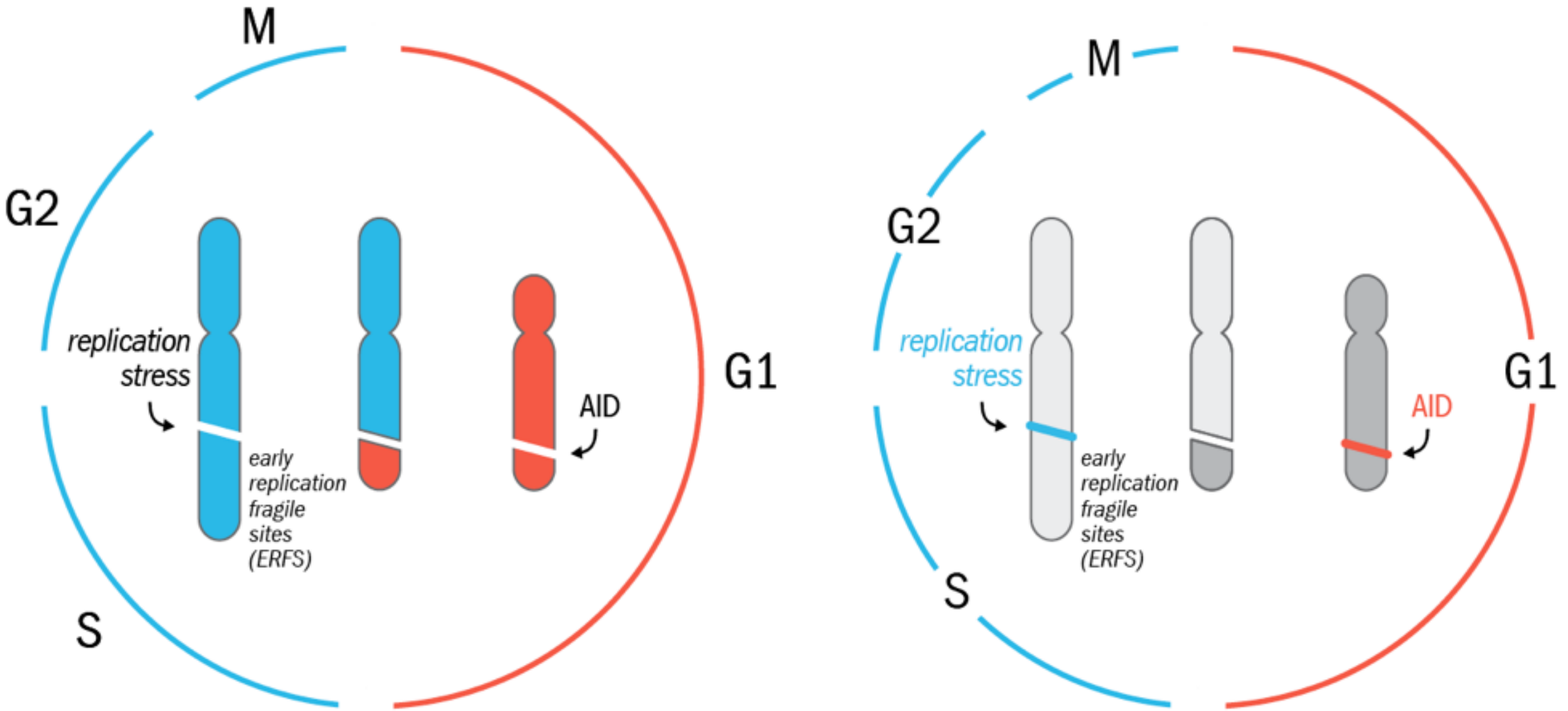


JH Barlow et al (2013) Identification of Early Replicating Fragile Sites that Contribute to Genome Instability. Cell 152:620–632.
redesign by M Krzywinski

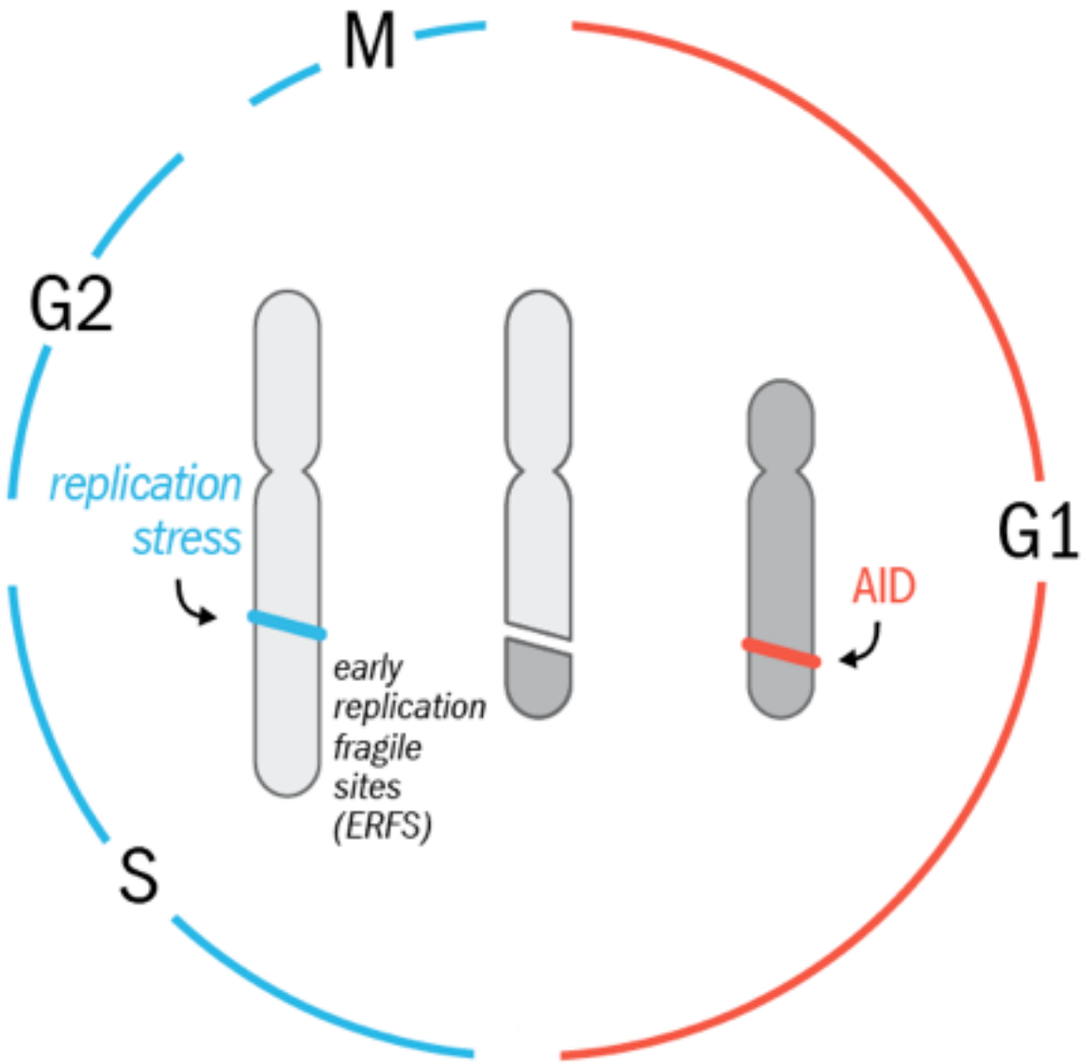


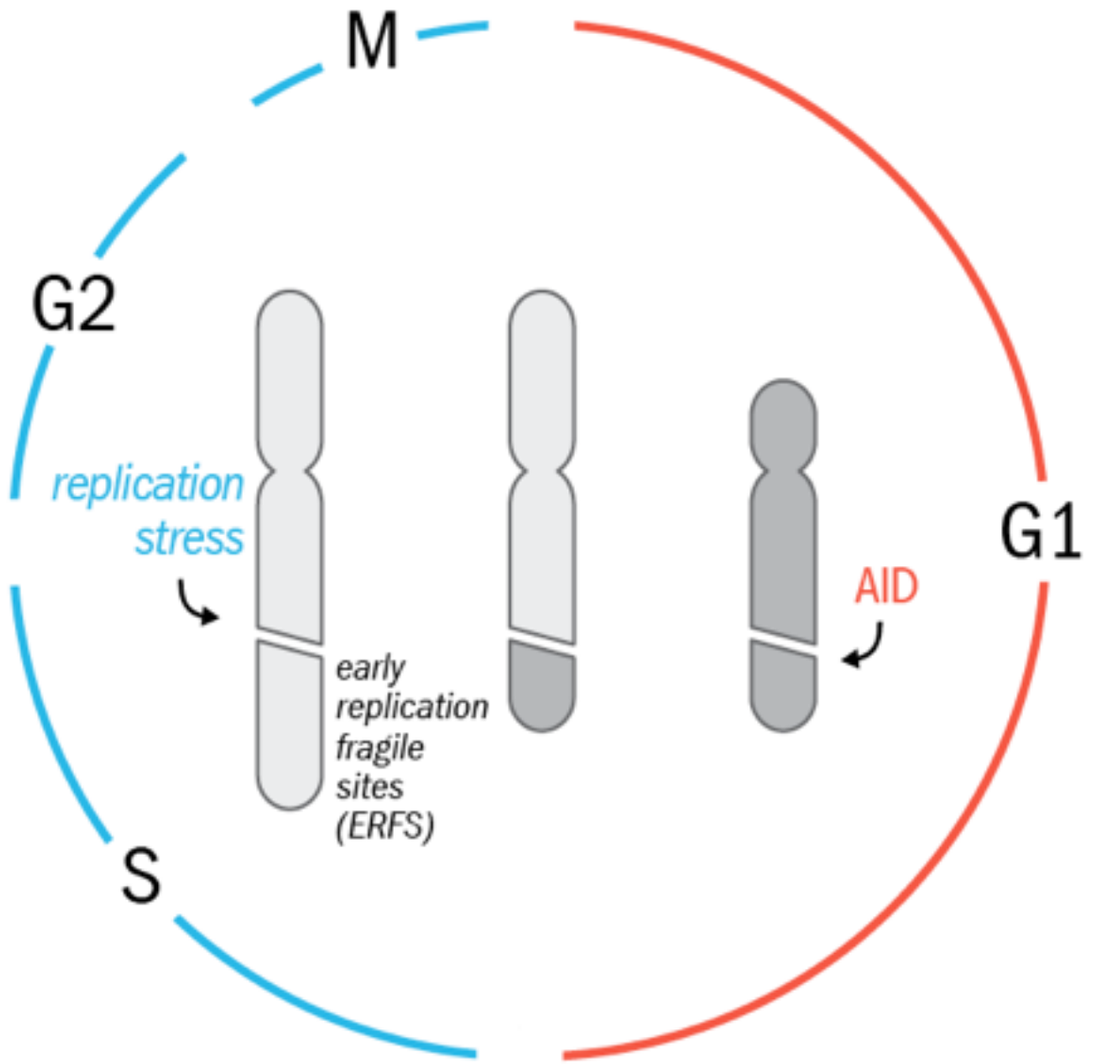


JH Barlow et al (2013) Identification of Early Replicating Fragile Sites that Contribute to Genome Instability. Cell 152:620–632.
redesign by M Krzywinski

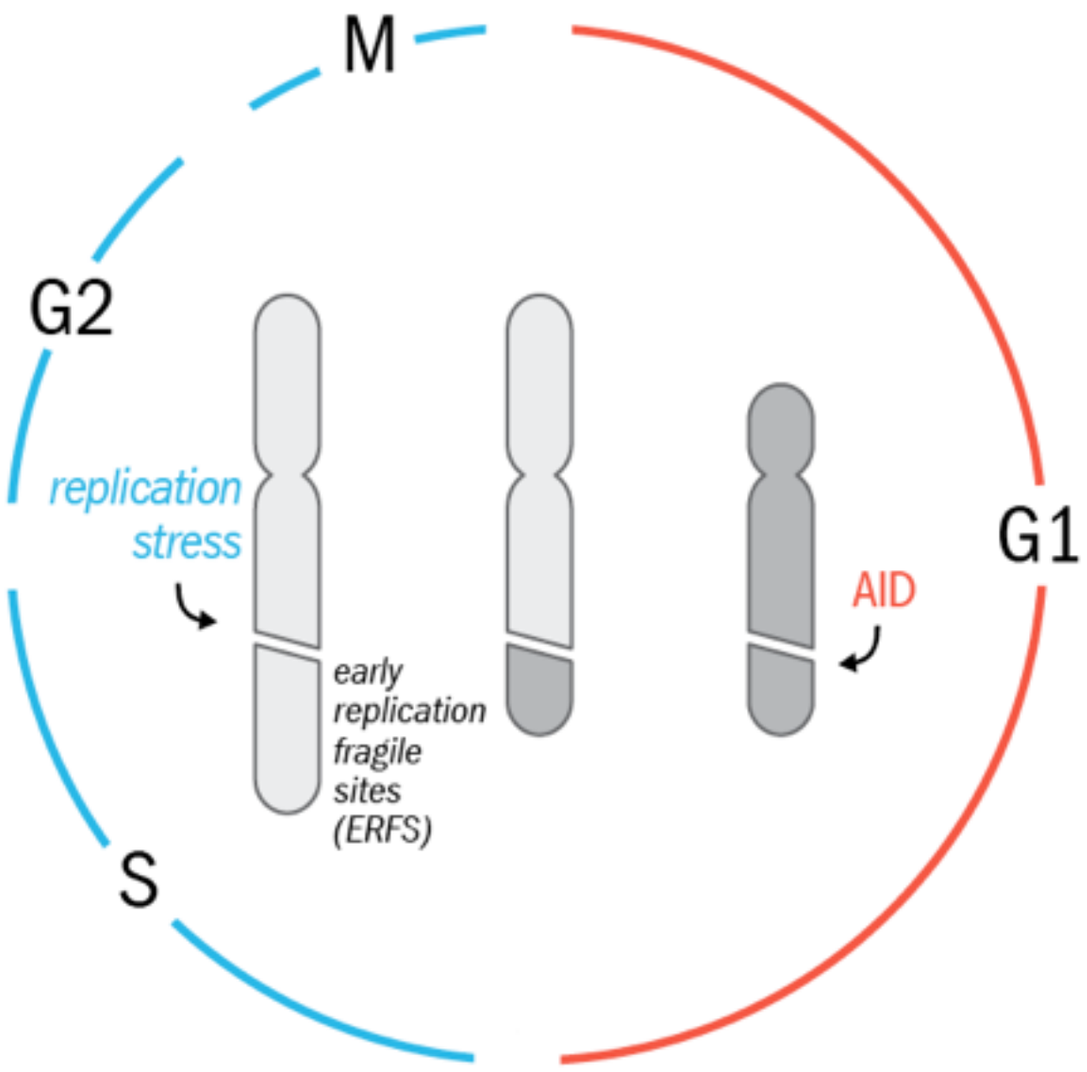


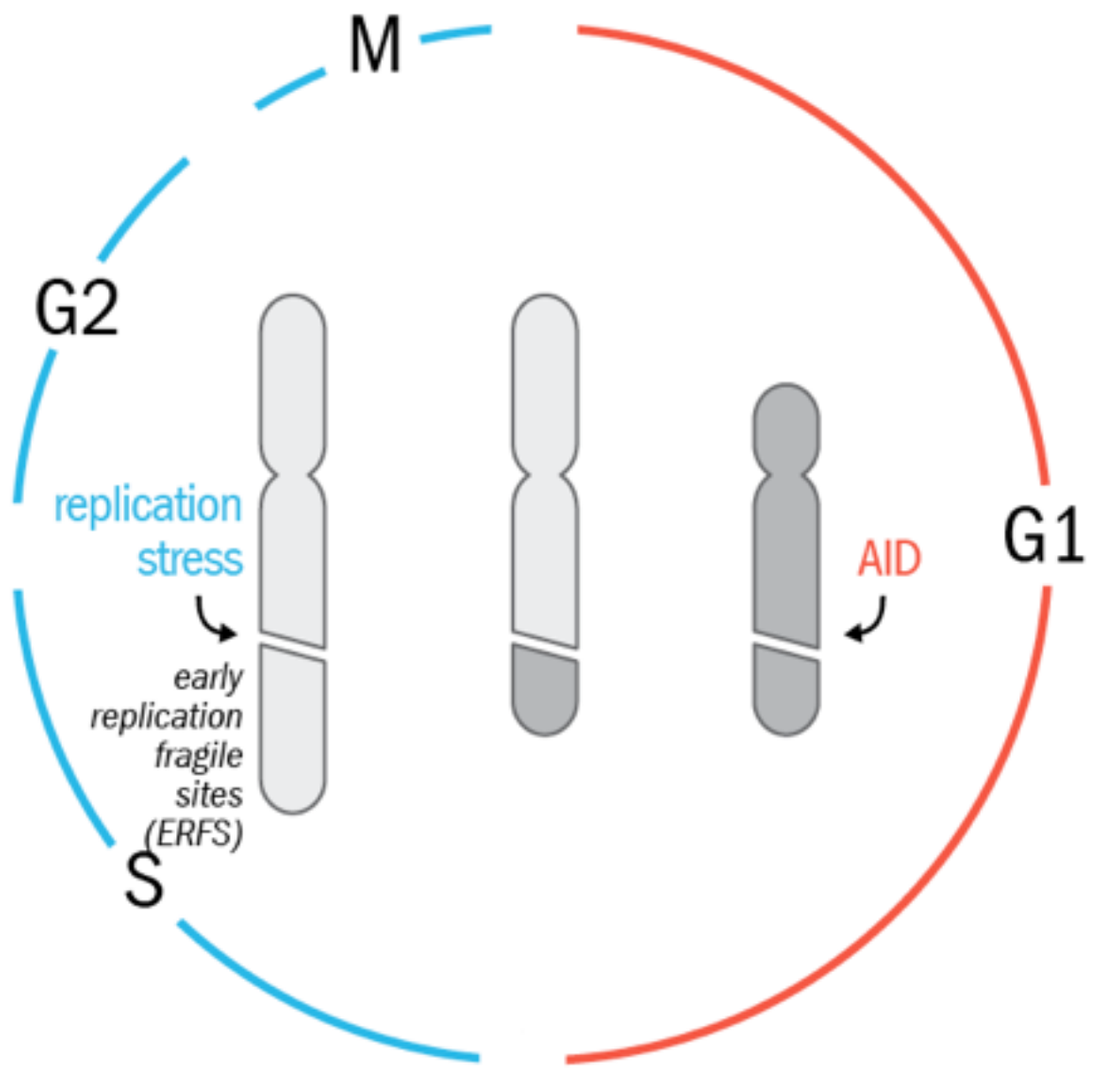
JH Barlow et al (2013) Identification of Early Replicating Fragile Sites that Contribute to Genome Instability. Cell 152:620–632.
redesign by M Krzywinski



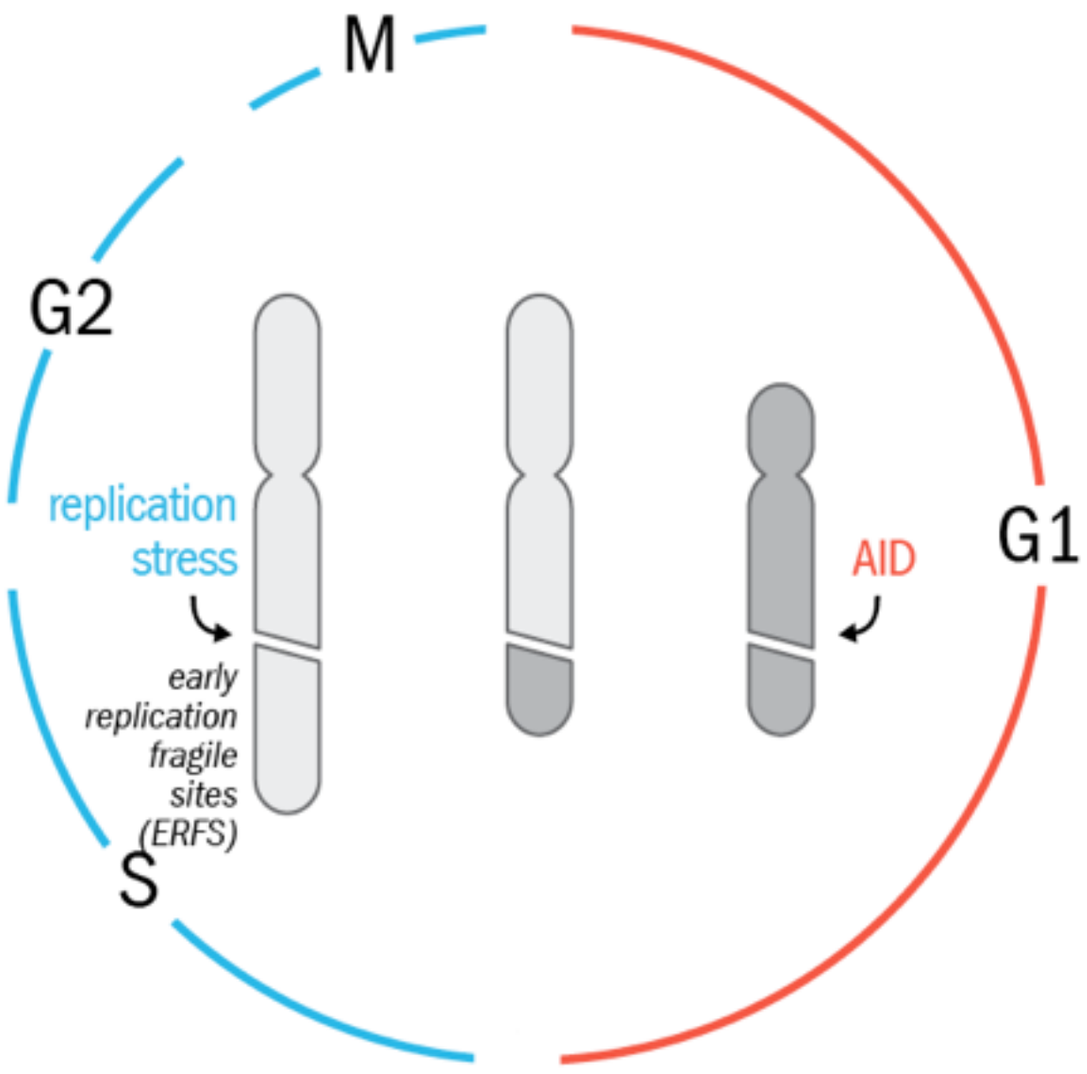


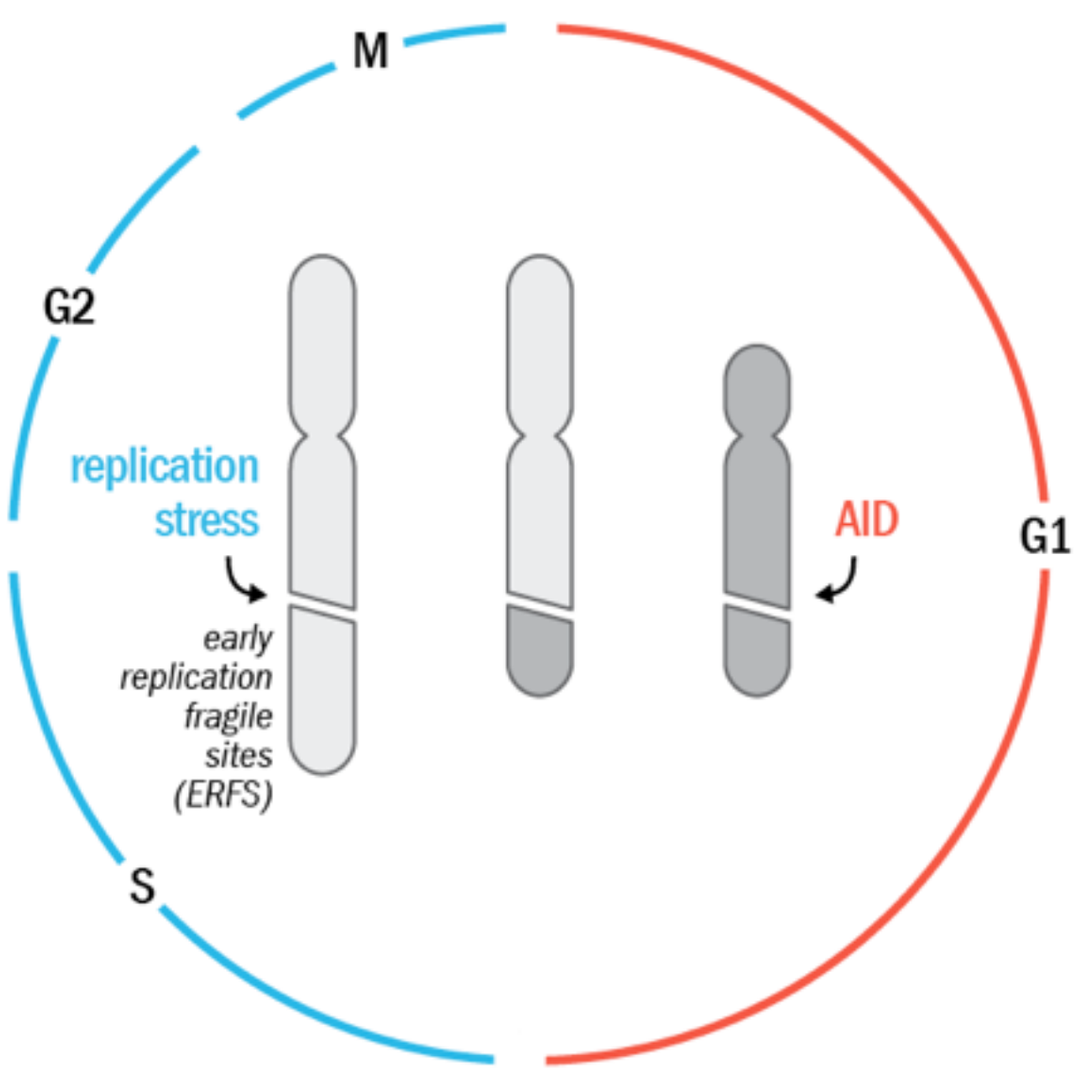
JH Barlow et al (2013) Identification of Early Replicating Fragile Sites that Contribute to Genome Instability. Cell 152:620–632.
redesign by M Krzywinski



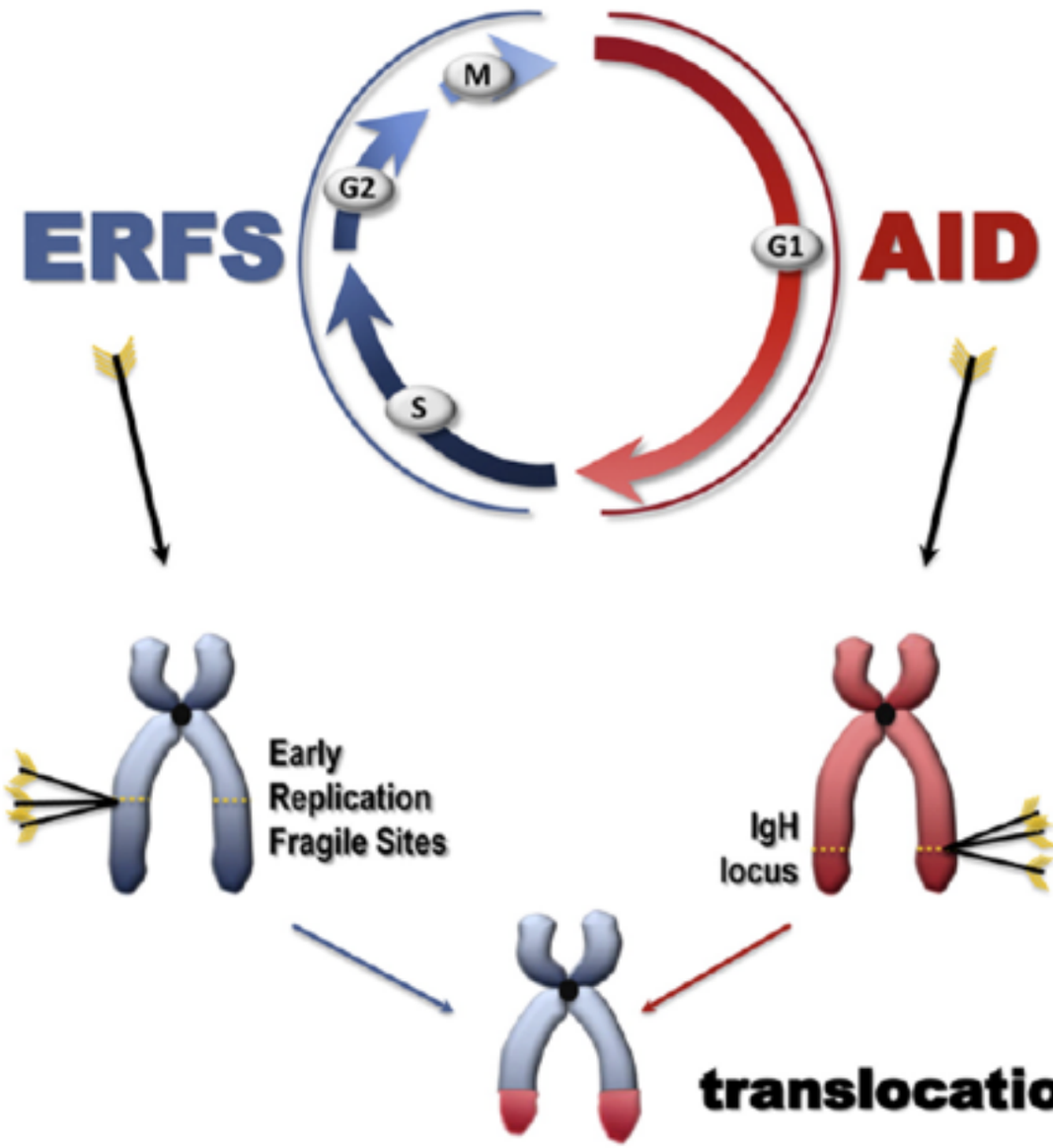


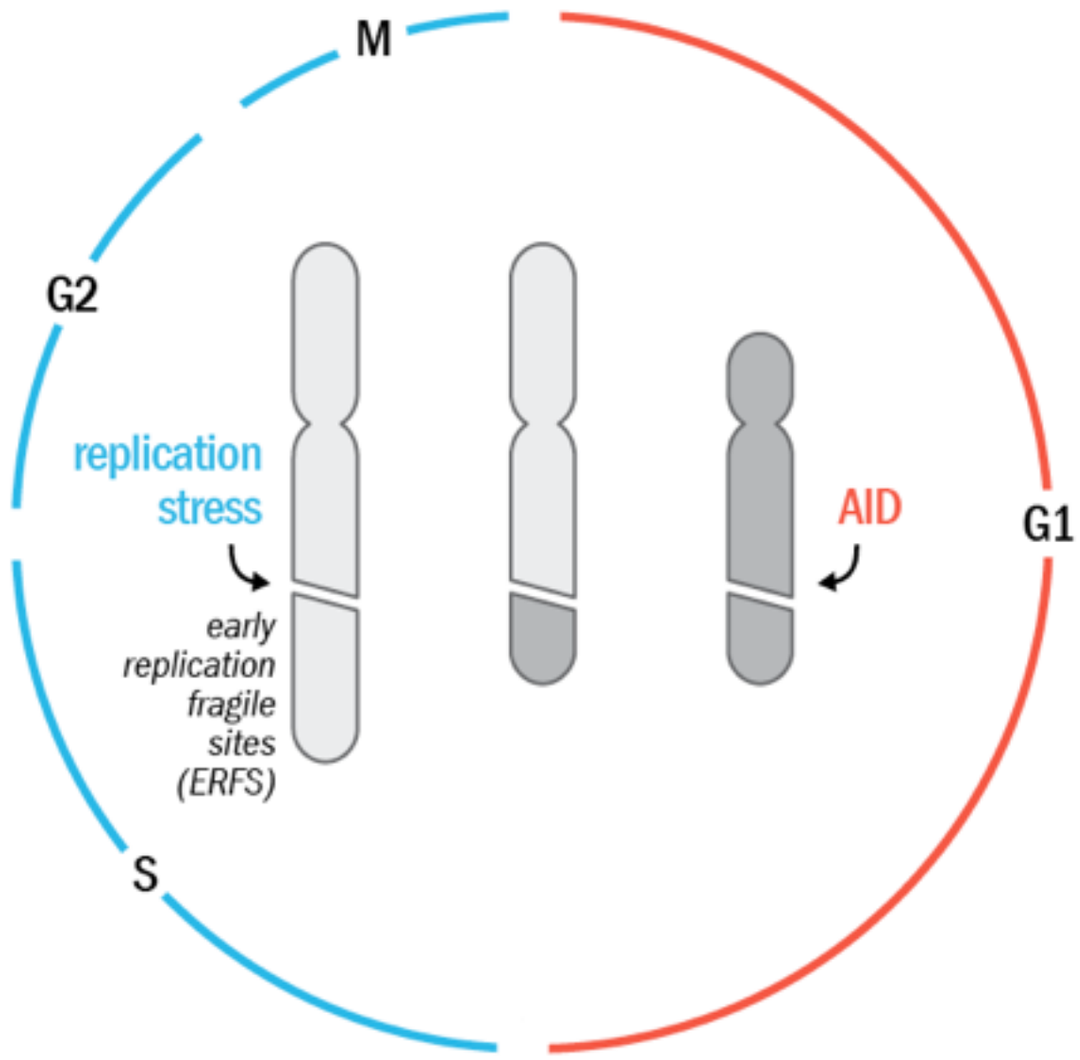
JH Barlow et al (2013) Identification of Early Replicating Fragile Sites that Contribute to Genome Instability. Cell 152:620–632.
redesign by M Krzywinski





JH Barlow et al (2013) Identification of Early Replicating Fragile Sites that Contribute to Genome Instability. Cell 152:620–632.
redesign by M Krzywinski





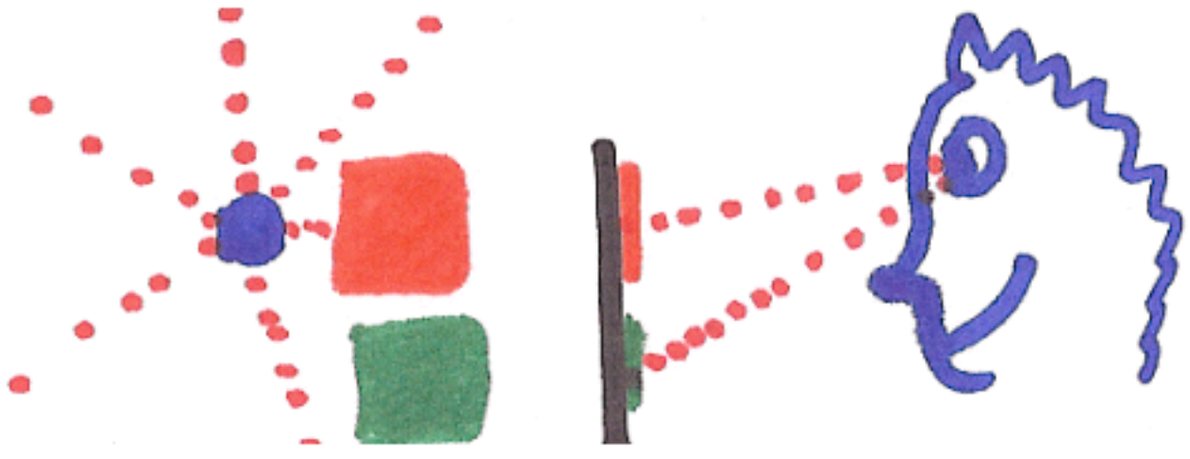
JH Barlow et al (2013) Identification of Early Replicating Fragile Sites that Contribute to Genome Instability. Cell 152:620–632.
redesign by M Krzywinski

3D FIGURES

What you should definitely not do

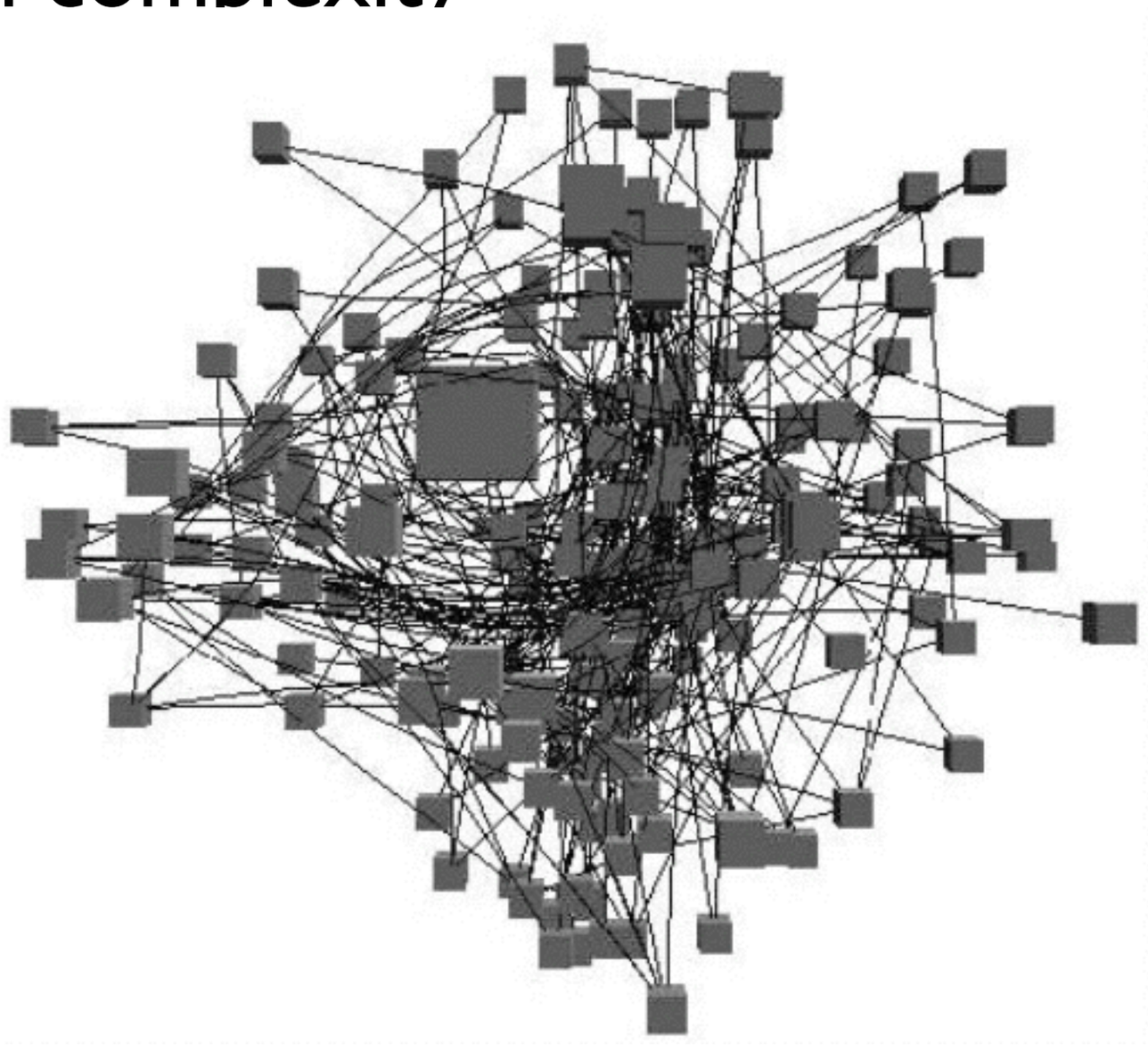
Dangers of depth

- rankings for planar spatial position, not depth!
- we don't really live in 3D: we **see** in 2.05D
  - up/down and sideways: image plane
    - acquire more info quickly from eye movements
  - away: depth into scene
    - only acquire more info from head/body motion



Dangers of depth: difficulties of 3D

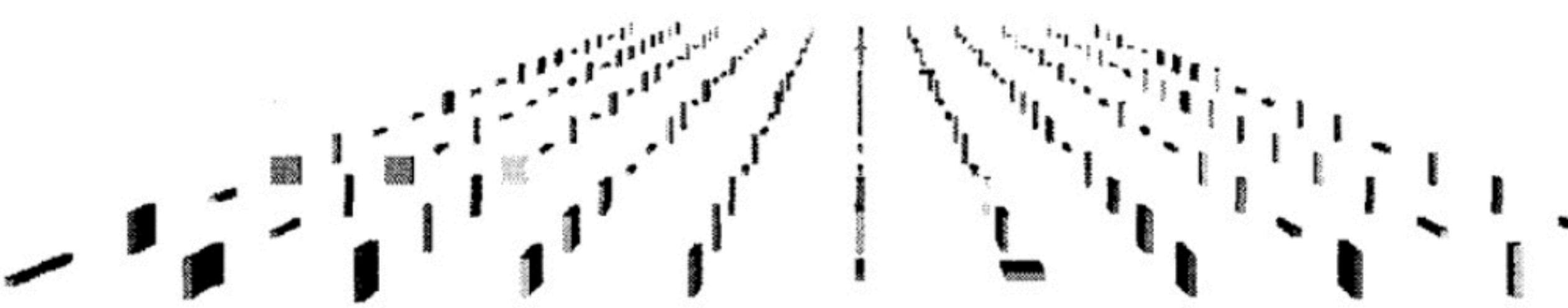
- occlusion
- interaction complexity



Distortion Viewing Techniques for 3D Data. Carpendale et al. InfoVis1996

Dangers of depth: difficulties of 3D

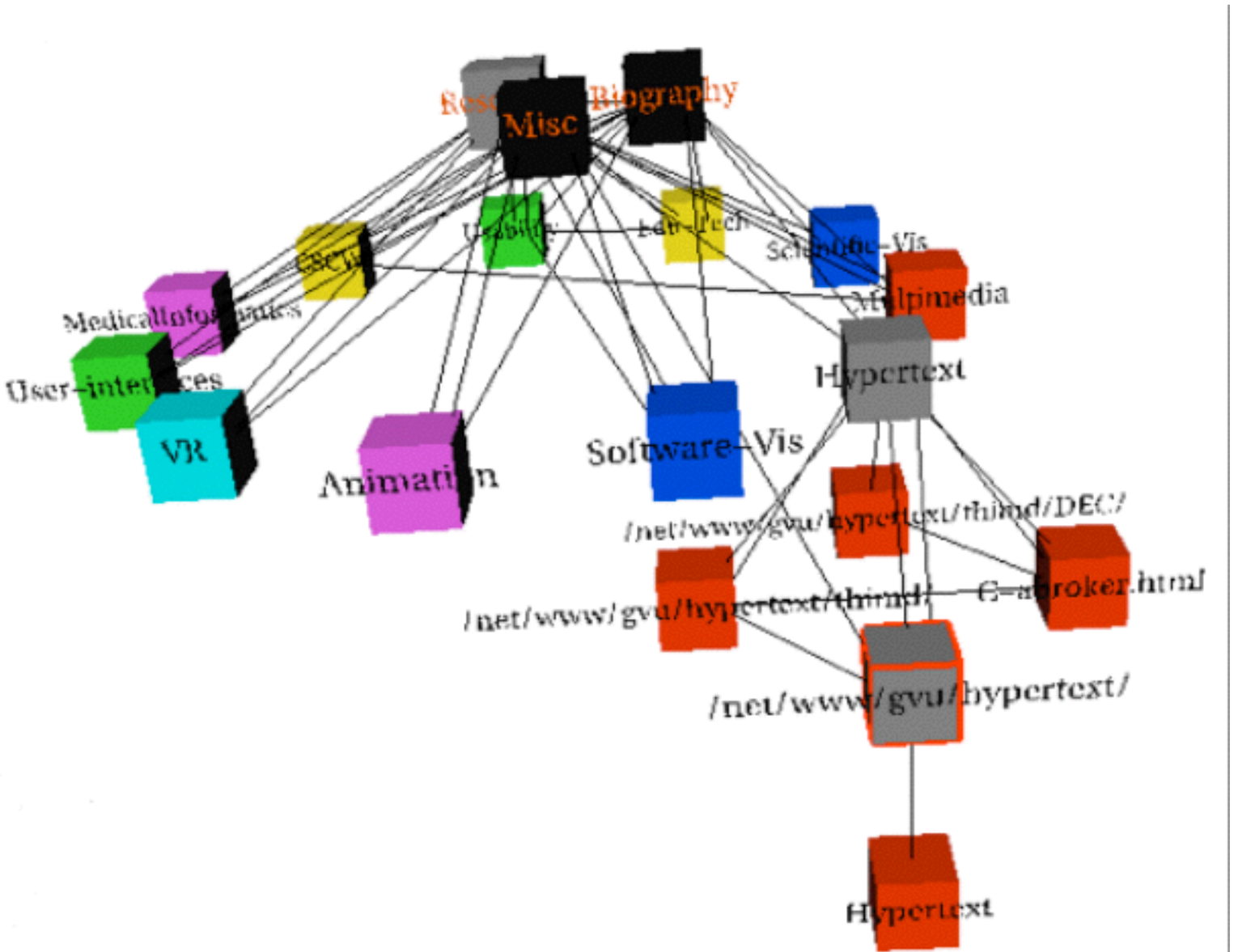
- perspective distortion
  - interferes with all size channel encodings
  - power of the plane is lost!



Visualizing the Results of Multimedia Web Search Engines.
Mukherjea, Hirata, and Hara. InfoVis 96

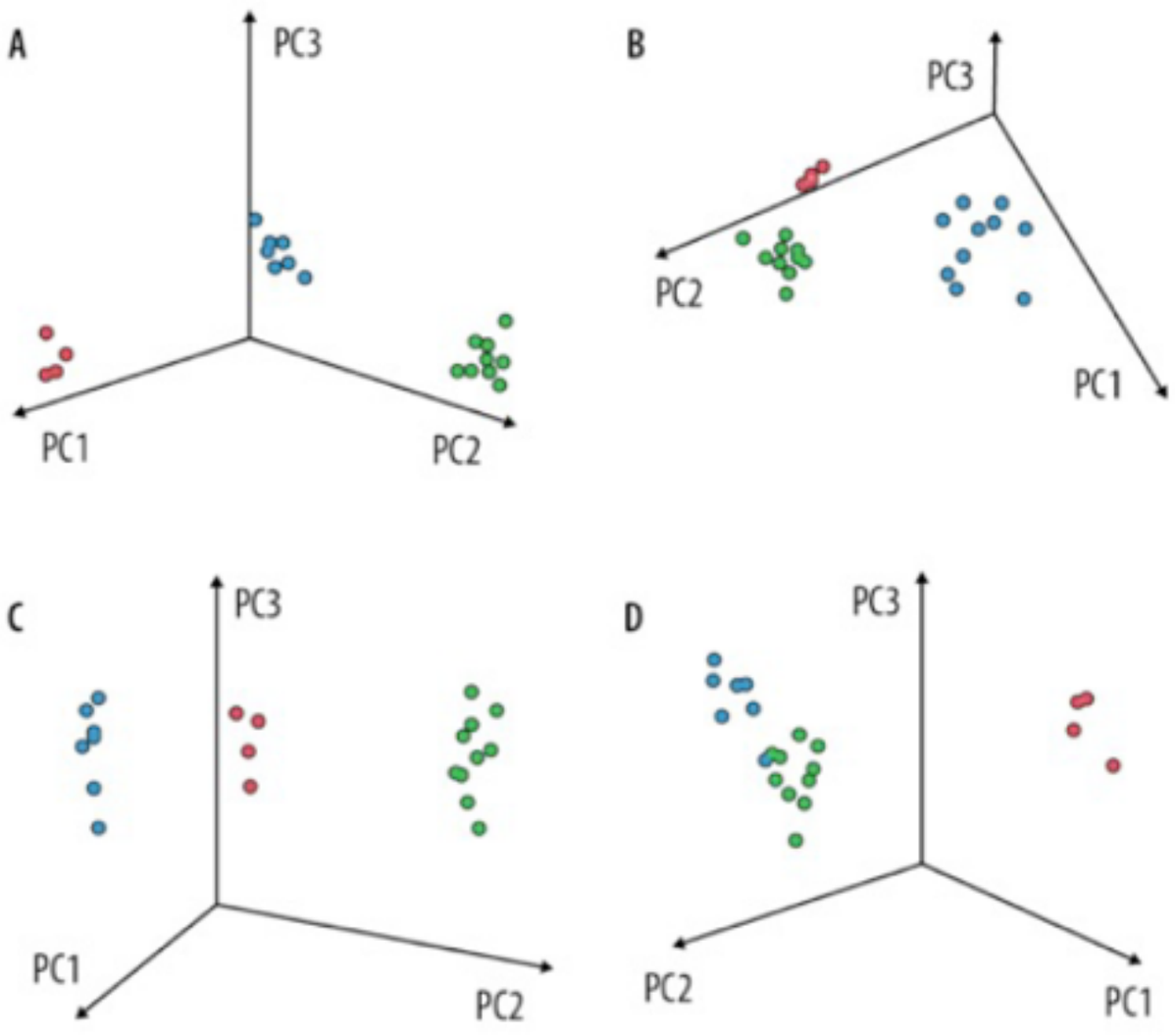
Dangers of depth: difficulties of 3D

- text legibility
  - far worse when tilted from image plane

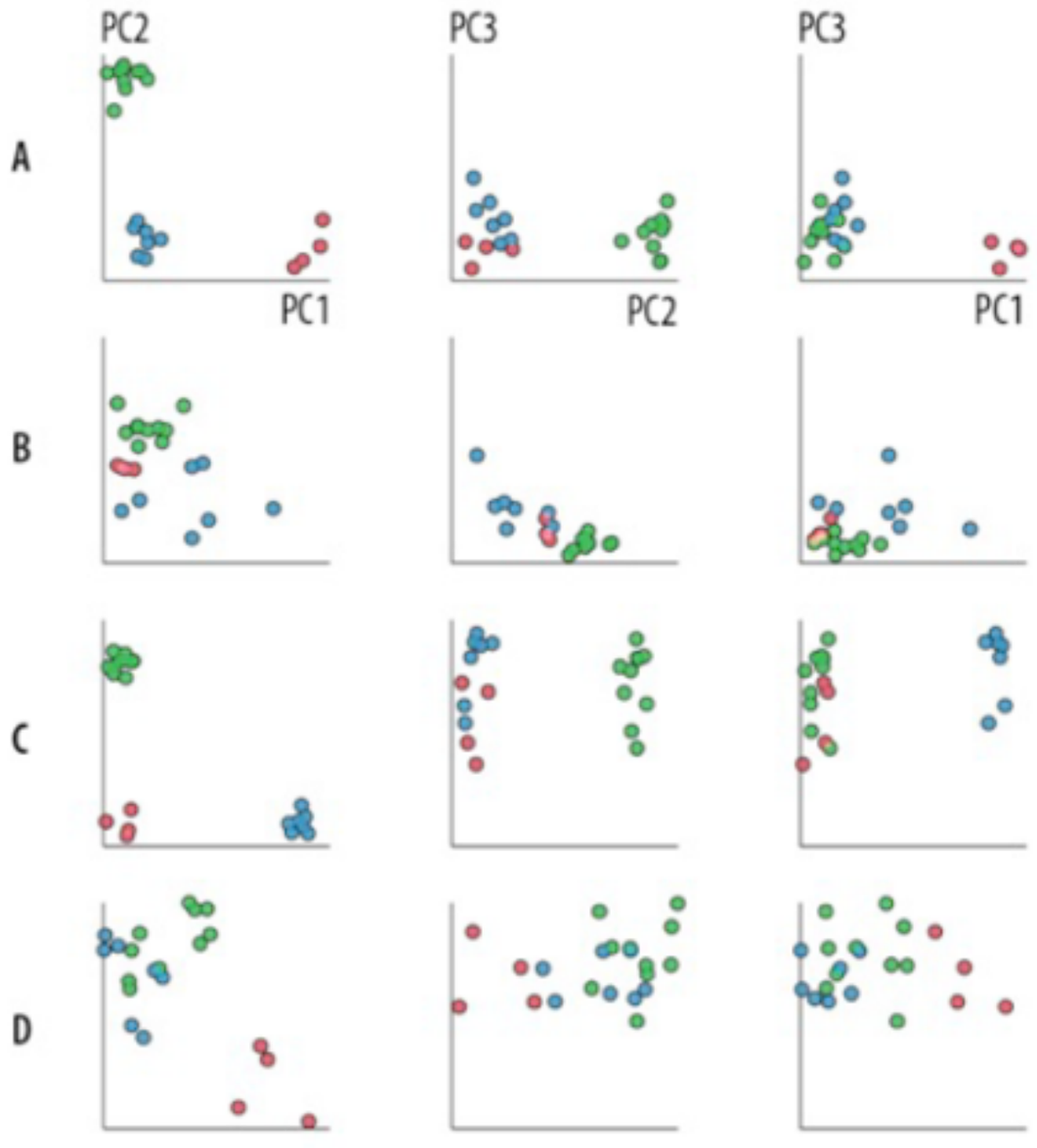


Confusing 3D graphs

confusing

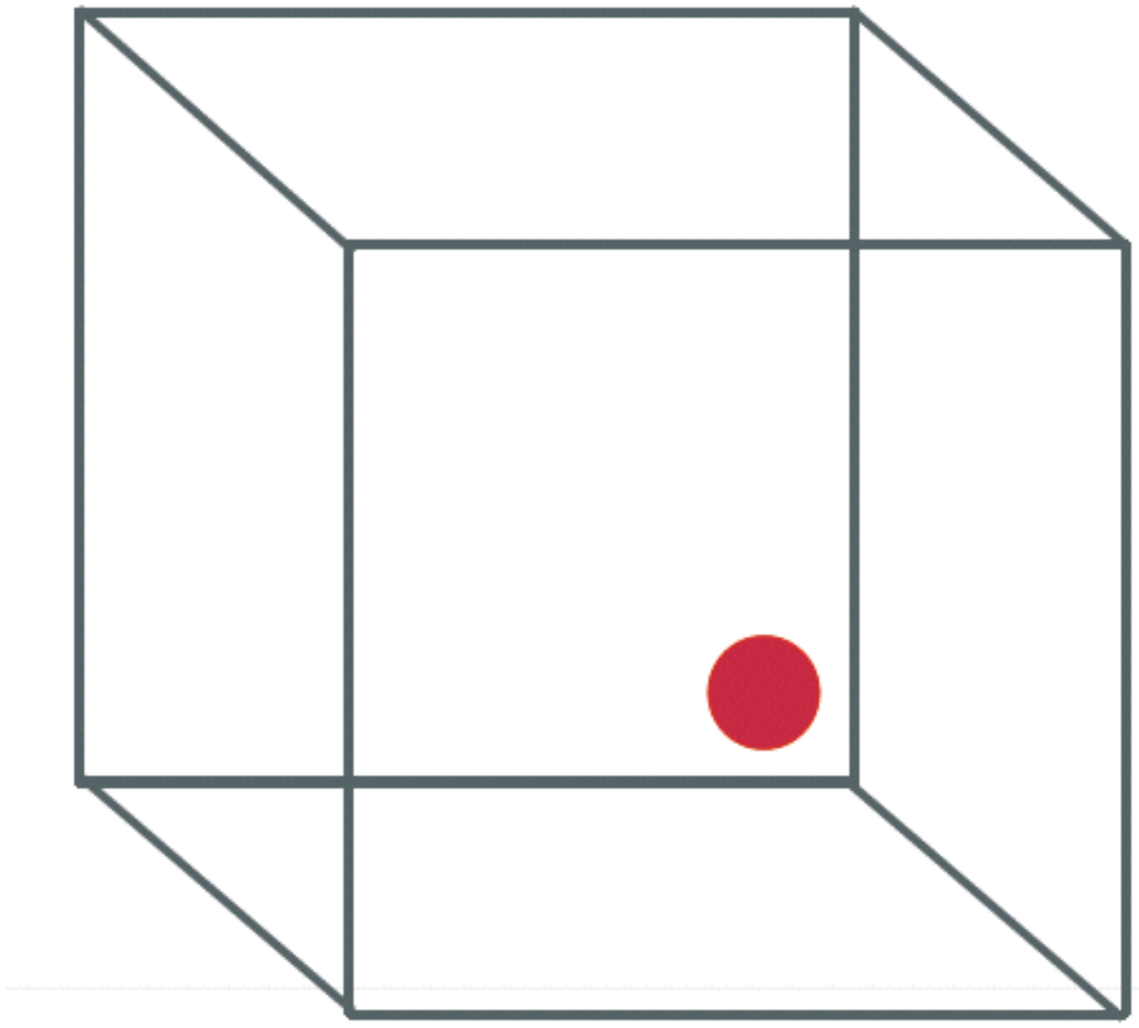


improved



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

Ambiguous Information: Position in 2.05D



VISUAL ABSTRACT CREATION

Identifying a deliverable message.

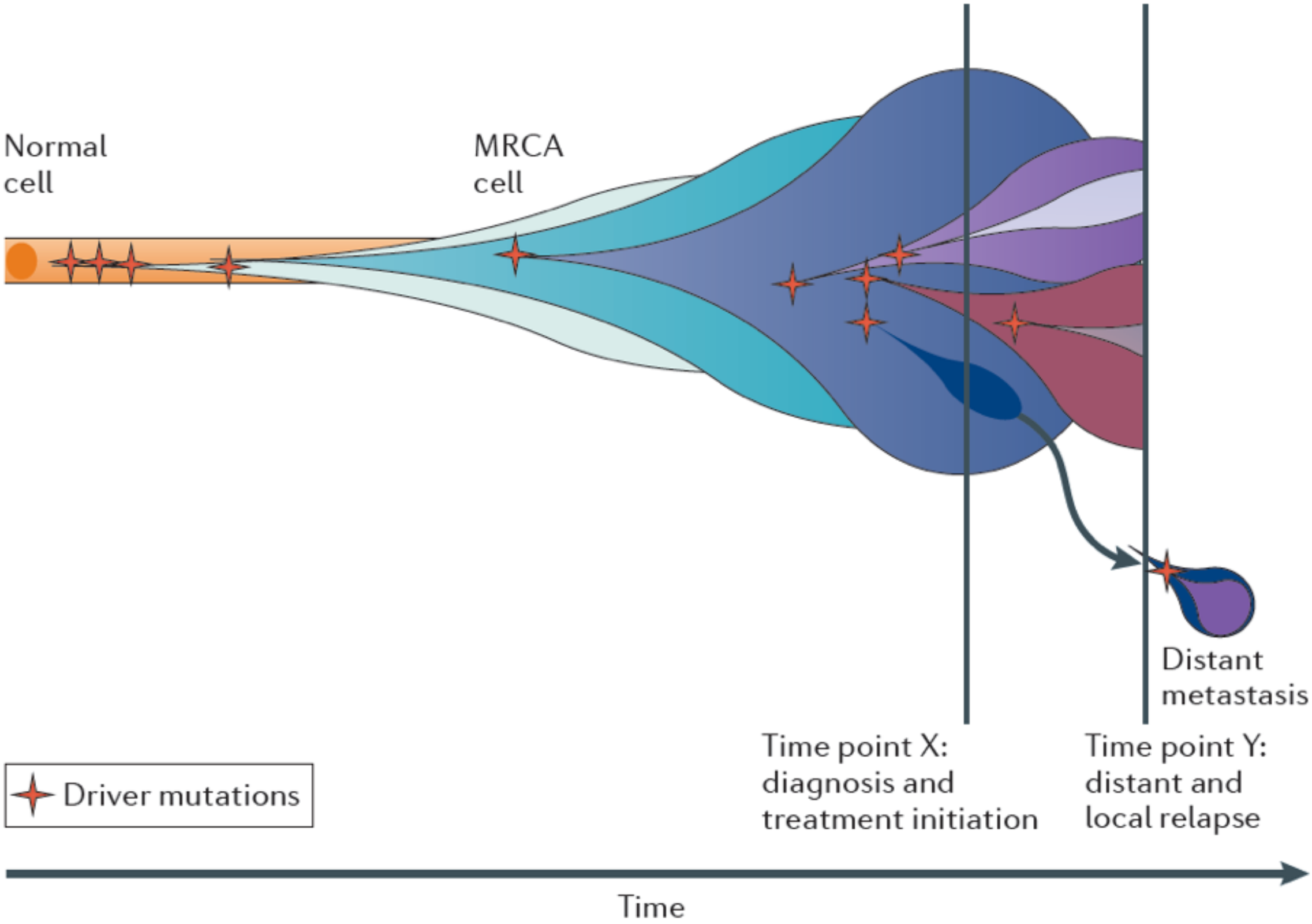
KEYWORDS

cancer
cancer genomics
tumor heterogeneity
next-generation sequencing
second-generation sequencing
third-generation sequencing
mutation discovery
whole genome sequencing
single molecule sequencing
single cell sequencing
personalized medicine

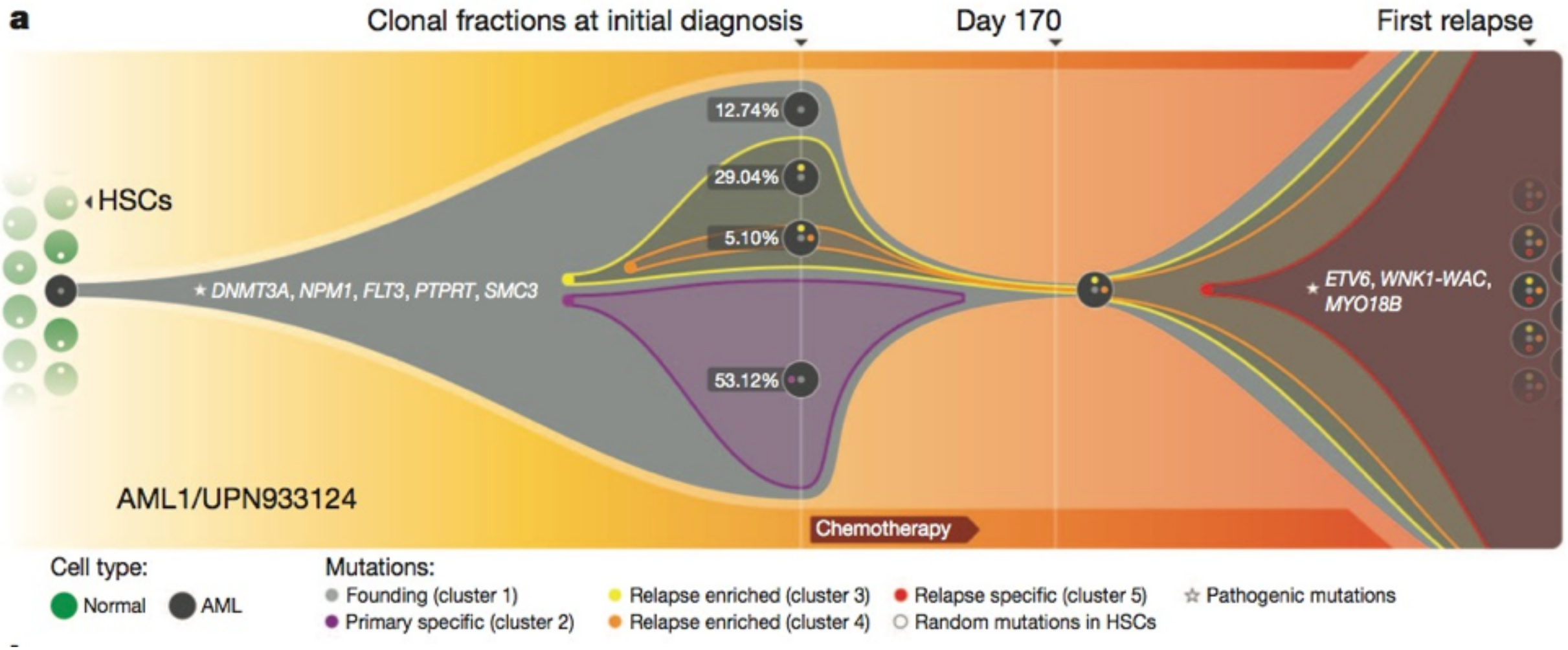
Chun HJE, Khattra J, Krzywinski M, Aparicio SA, Marra MA (Dellaire GD, Berman JN, Arceci RJ. editors). Cancer Genomics. 1st ed. Elsevier; 2013.

ELSEVIER FIGURE RESTRICTION

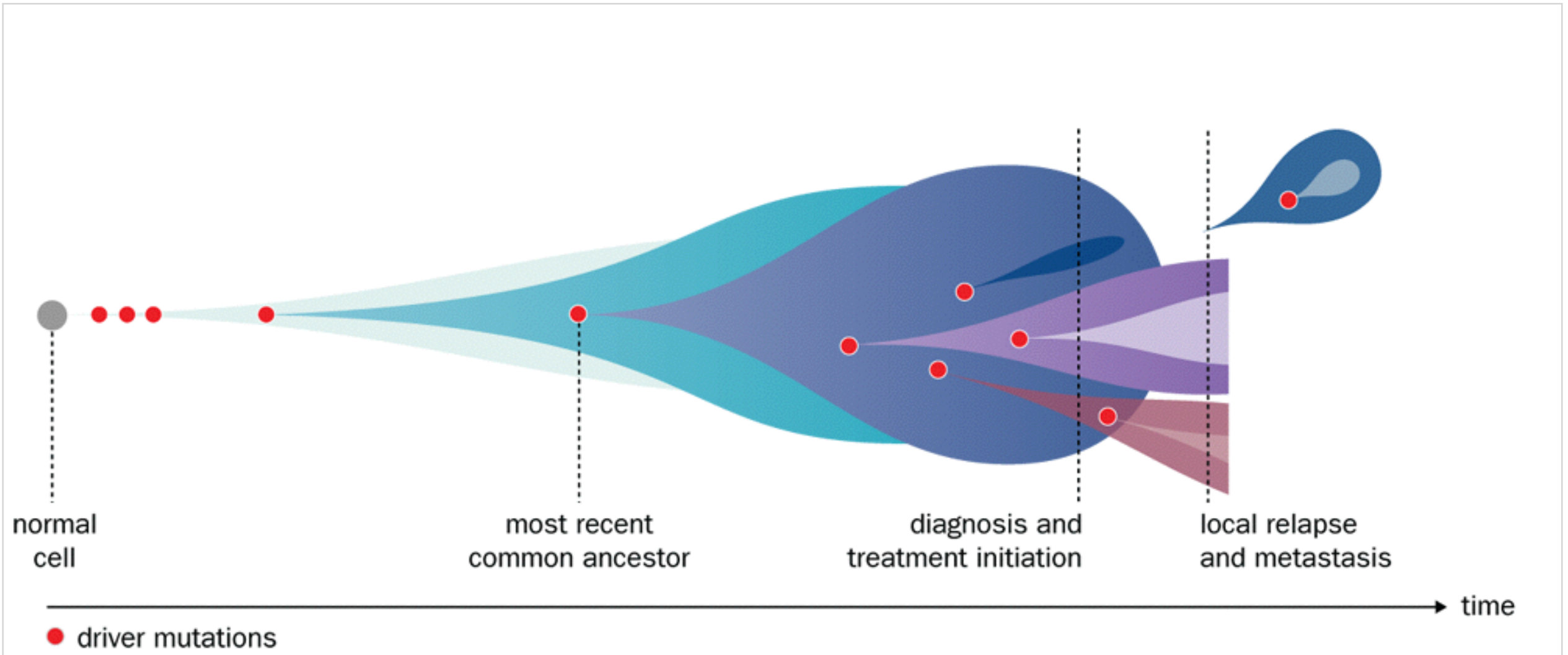
Please note that your image will be scaled proportionally to fit in the available window on ScienceDirect, **a 500 by 200 pixel rectangle**.



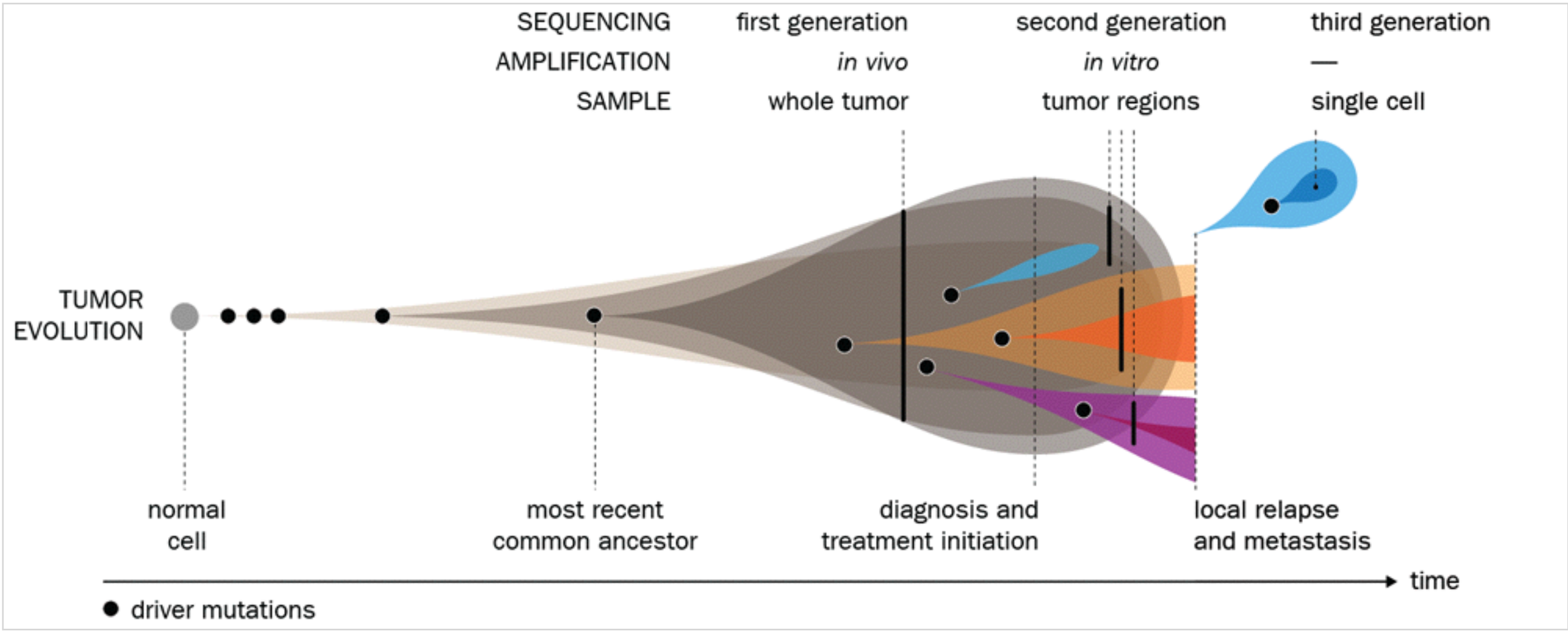
Yates LR, Campbell PJ. Evolution of the cancer genome. Nat Rev Genet 2012 Nov;13(11):795-806.



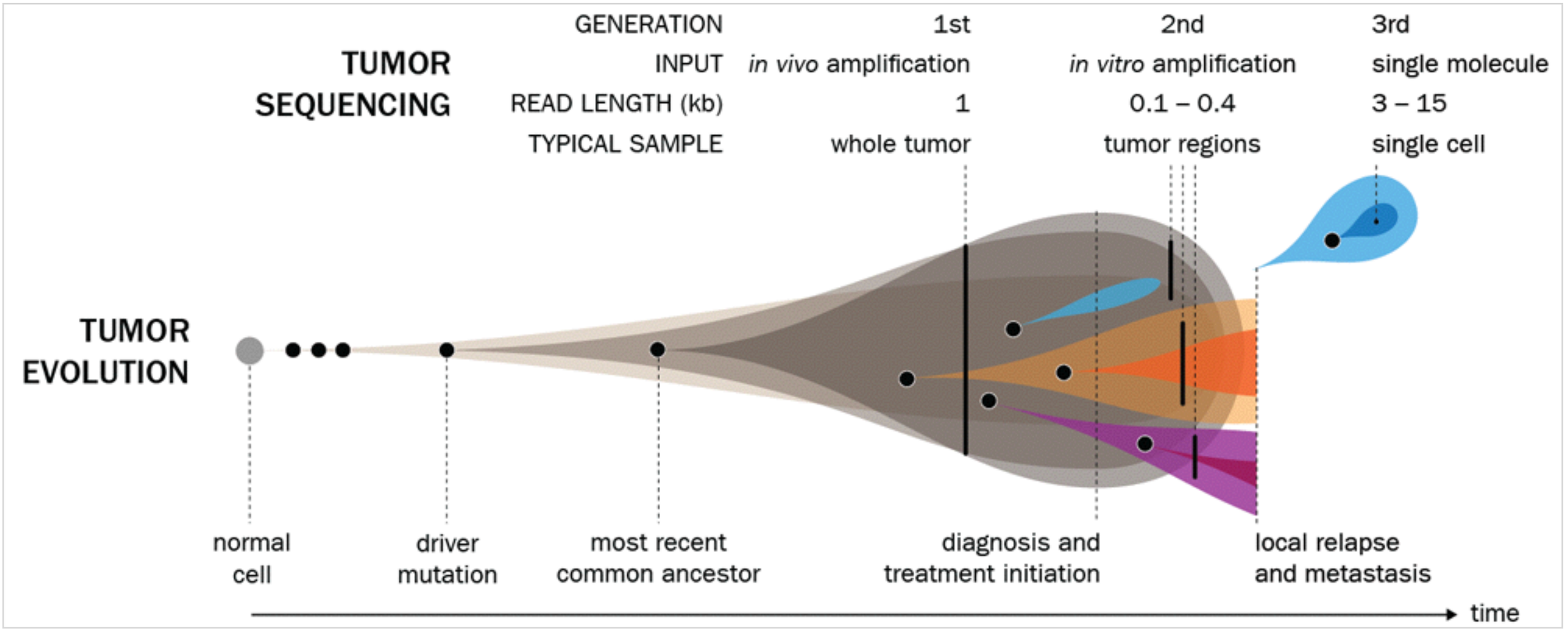
Ding L, Ley TJ, Larson DE, Miller CA, Koboldt DC, Welch JS, et al. Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. Nature. 2012 Jan 26;481(7382):506-10.



Chun HJE, Khattra J, Krzywinski M, Aparicio SA, Marra MA (Dellaire GD, Berman JN, Arceci RJ. editors). Cancer Genomics. 1st ed. Elsevier; 2013.

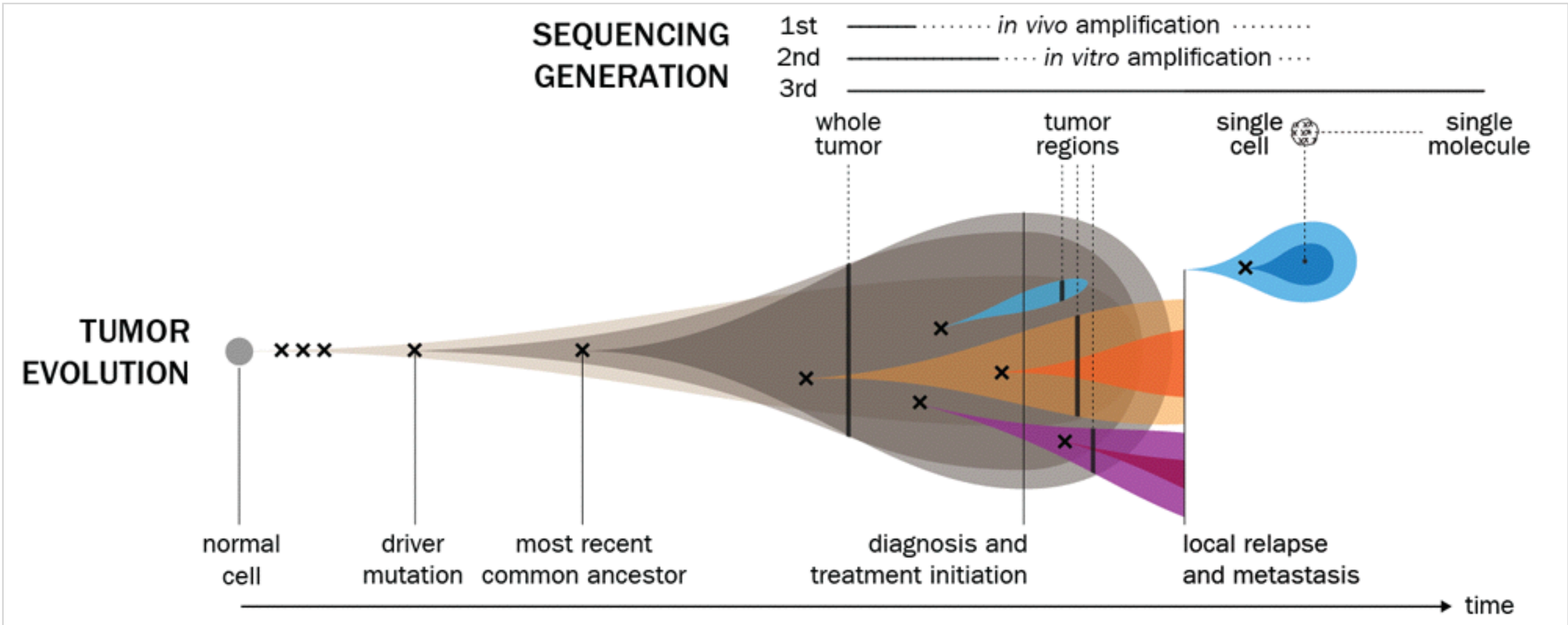


Chun HJE, Khattra J, Krzywinski M, Aparicio SA, Marra MA (Dellaire GD, Berman JN, Arceci RJ. editors). Cancer Genomics. 1st ed. Elsevier; 2013.
Early version of visual abstract.



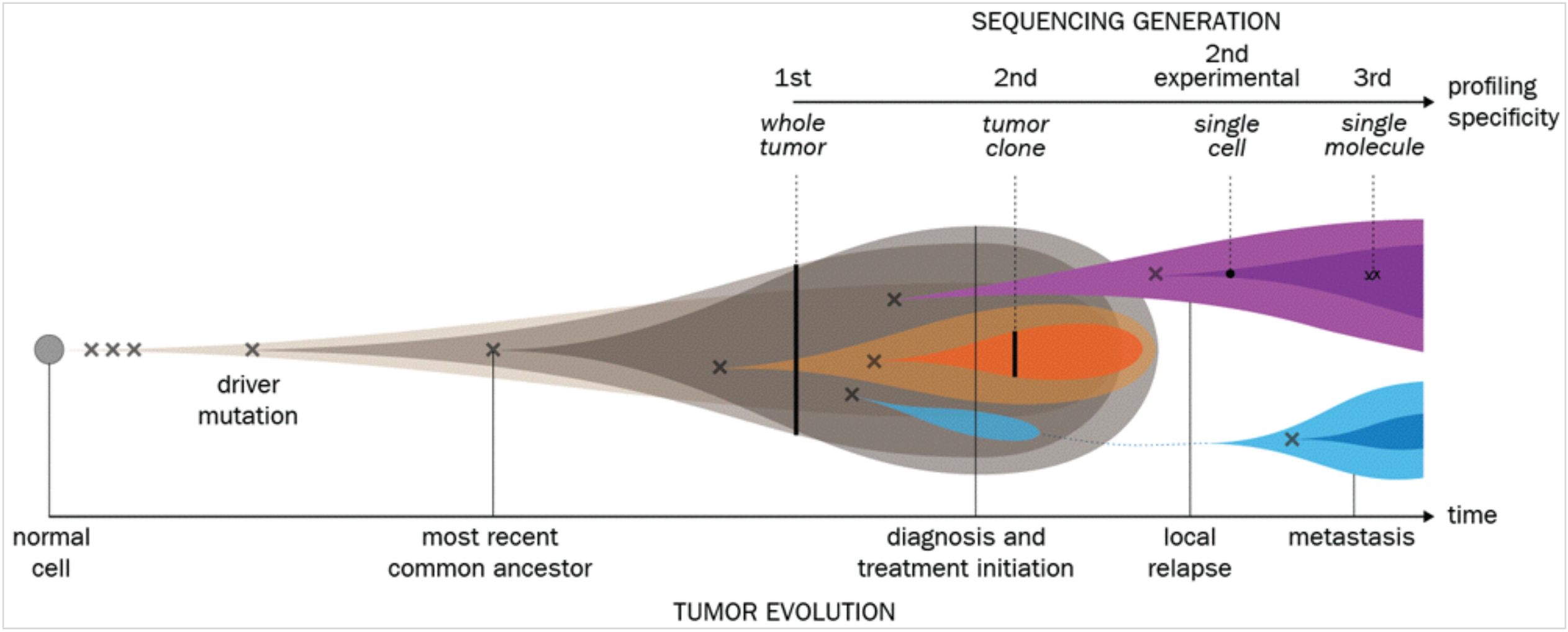
Chun HJE, Khattra J, Krzywinski M, Aparicio SA, Marra MA (Dellaire GD, Berman JN, Arceci RJ. editors). Cancer Genomics. 1st ed. Elsevier; 2013.
Early version of visual abstract.

VIZBI 2013



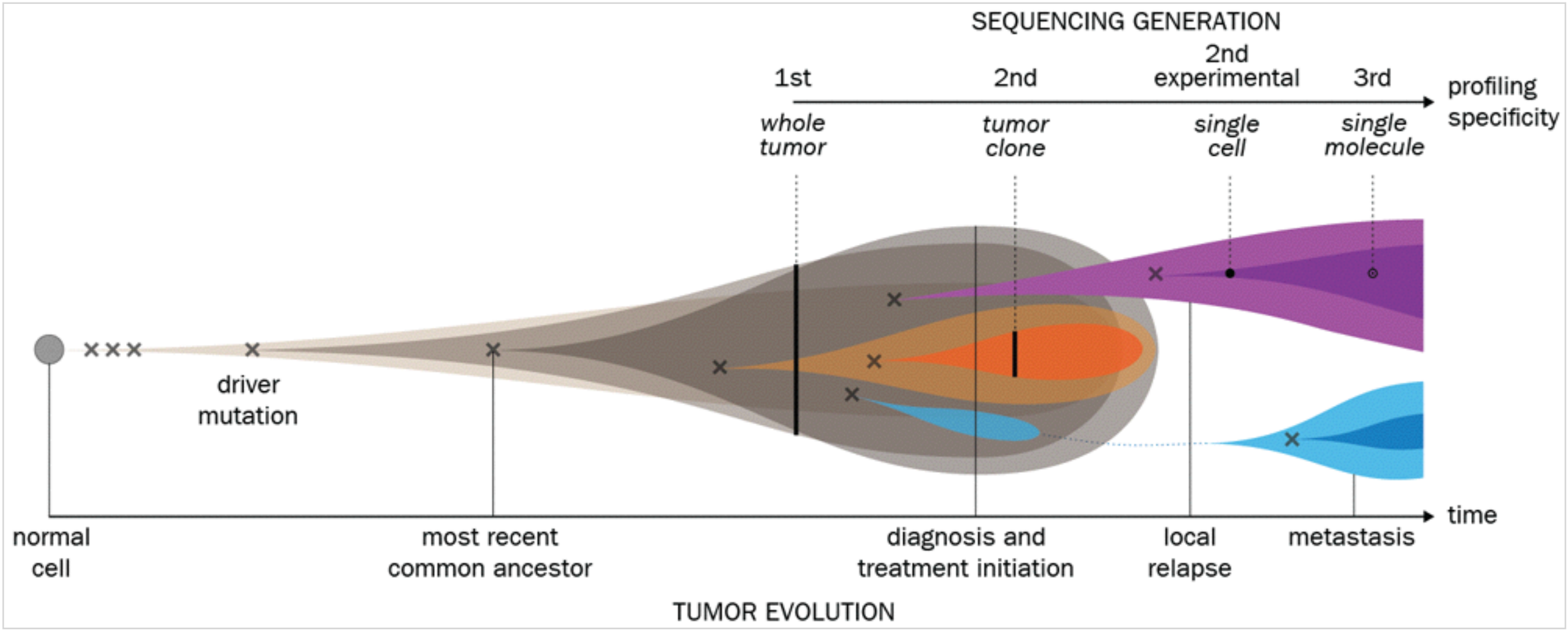
Chun HJE, Khattra J, Krzywinski M, Aparicio SA, Marra MA (Dellaire GD, Berman JN, Arceci RJ. editors). Cancer Genomics. 1st ed. Elsevier; 2013.
Early version of visual abstract.

VIZBI 2013



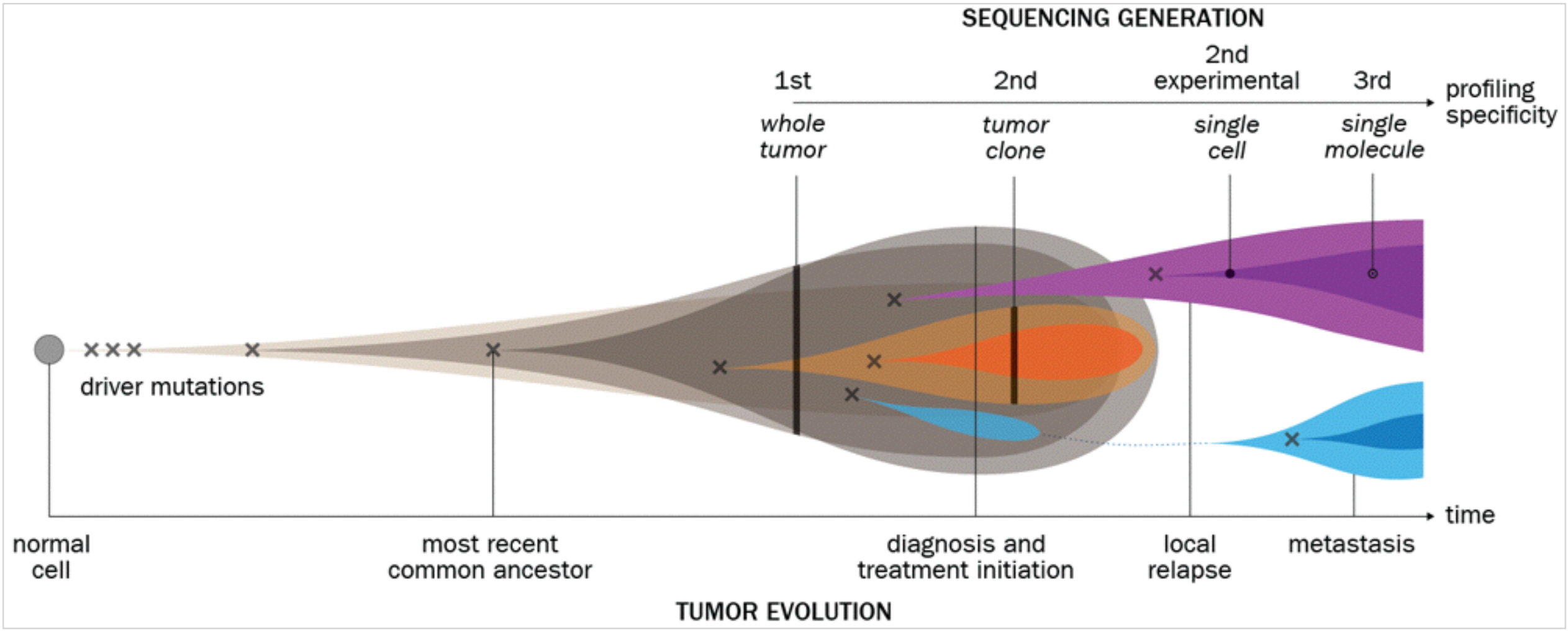
Chun HJE, Khattra J, Krzywinski M, Aparicio SA, Marra MA (Dellaire GD, Berman JN, Arceci RJ. editors). Cancer Genomics. 1st ed. Elsevier; 2013.
Early version of visual abstract.

VIZBI 2013



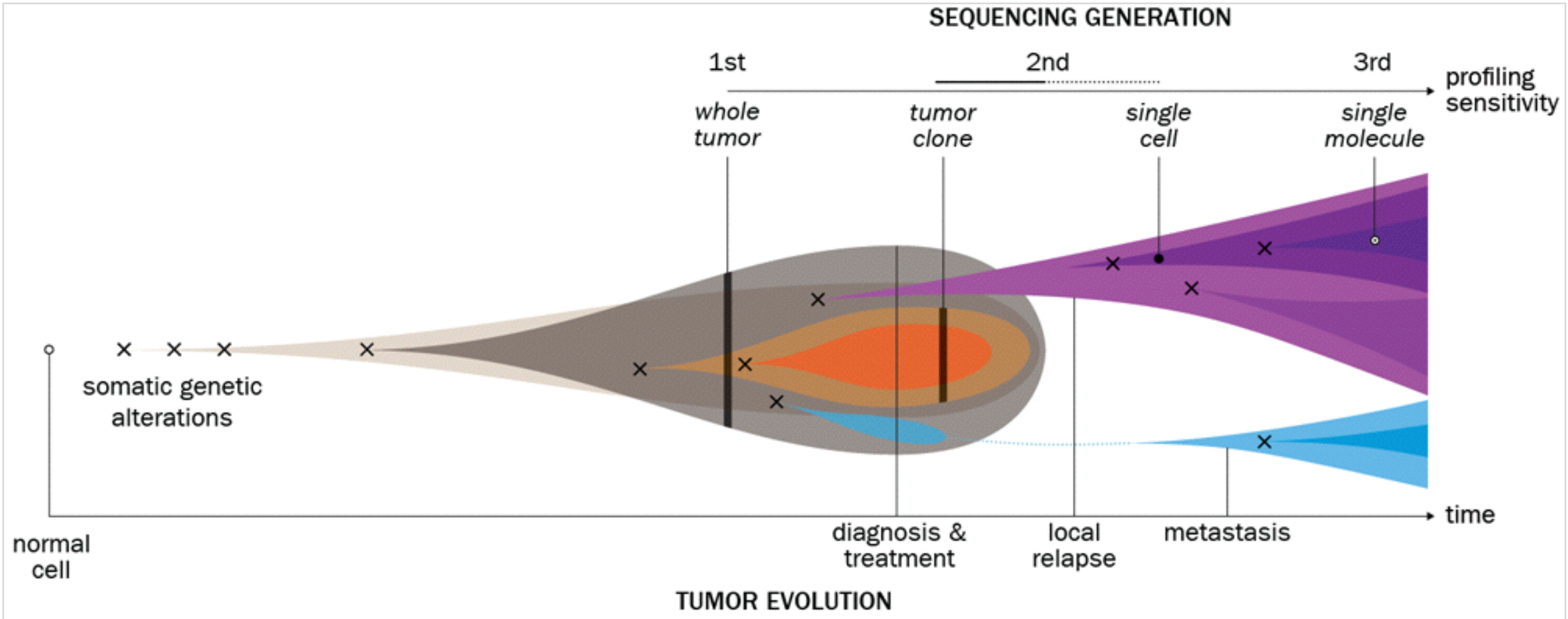
Chun HJE, Khattra J, Krzywinski M, Aparicio SA, Marra MA (Dellaire GD, Berman JN, Arceci RJ. editors). Cancer Genomics. 1st ed. Elsevier; 2013.
Early version of visual abstract.

VIZBI 2013

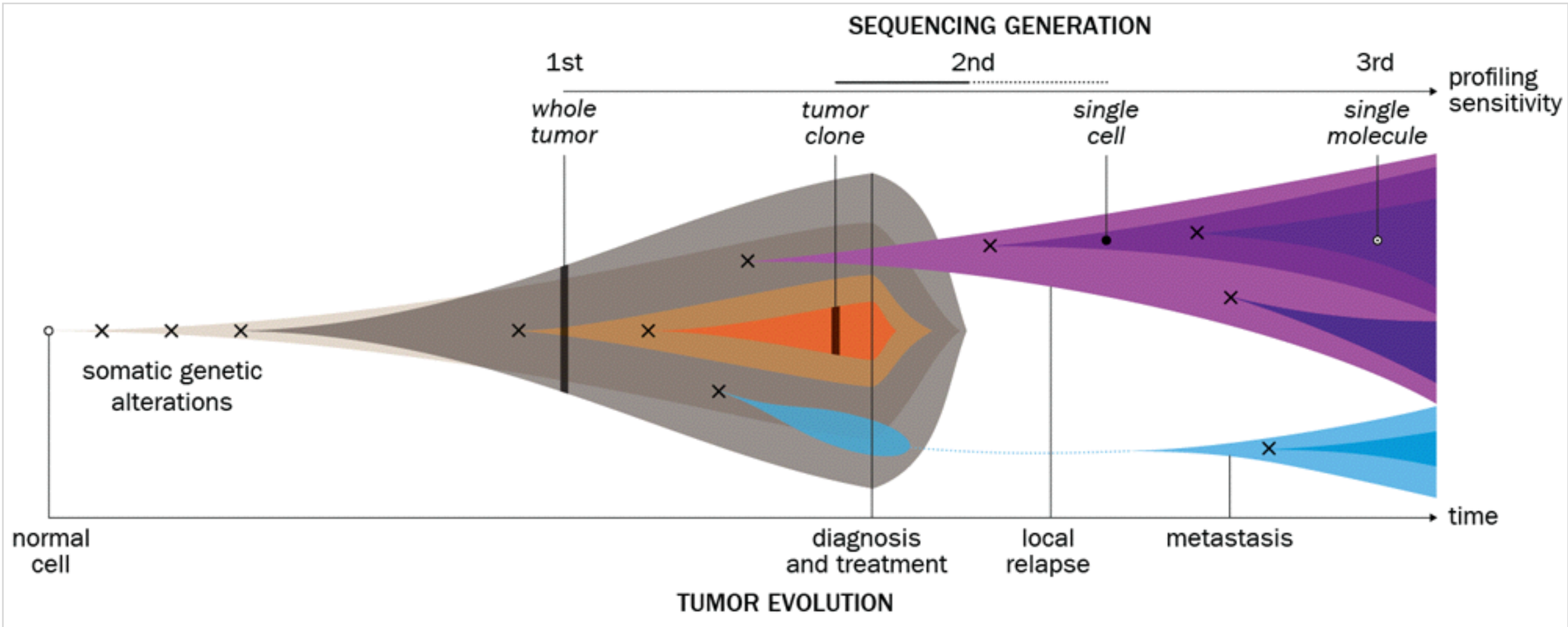


Chun HJE, Khattra J, Krzywinski M, Aparicio SA, Marra MA (Dellaire GD, Berman JN, Arceci RJ. editors). Cancer Genomics. 1st ed. Elsevier; 2013.
Early version of visual abstract.

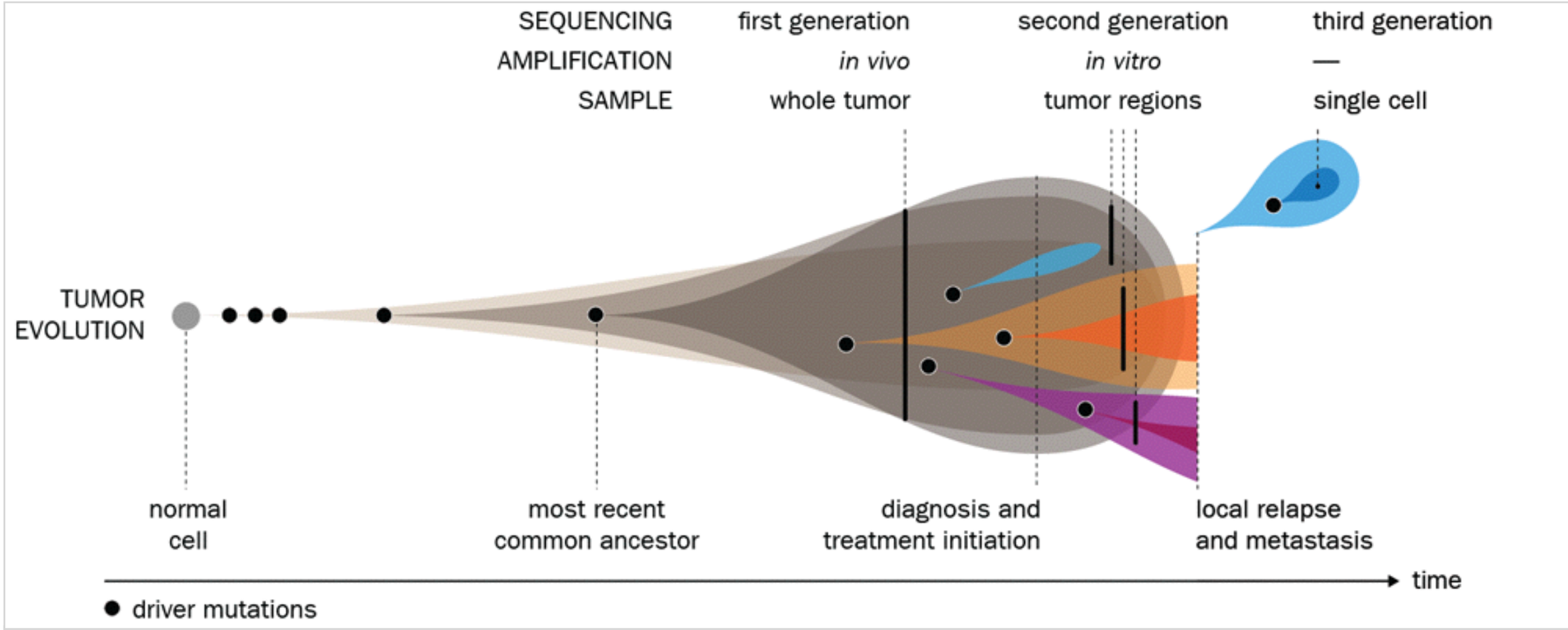
VIZBI 2013

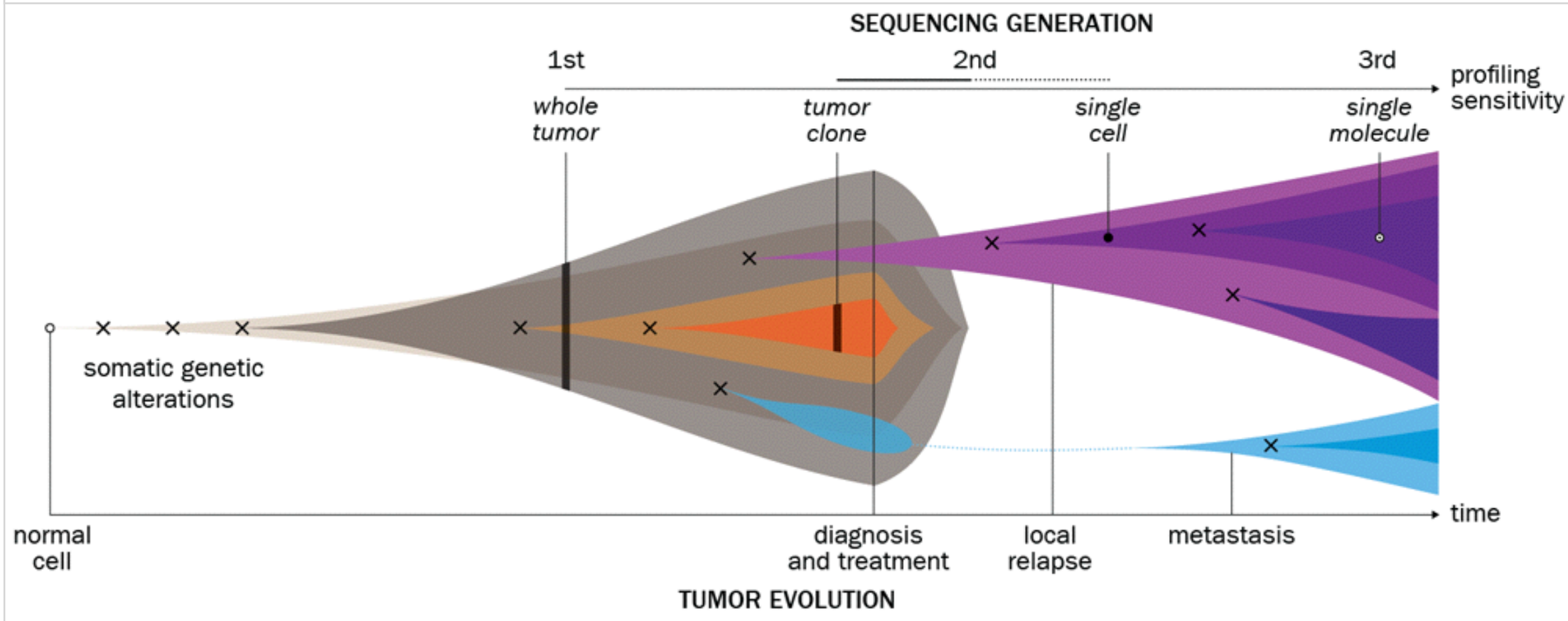


Chun HJE, Khattra J, Krzywinski M, Aparicio SA, Marra MA (Dellaire GD, Berman JN, Arceci RJ. editors). Cancer Genomics. 1st ed. Elsevier; 2013.
Early version of visual abstract.



Chun HJE, Khattra J, Krzywinski M, Aparicio SA, Marra MA (Dellaire GD, Berman JN, Arceci RJ. editors). Cancer Genomics. 1st ed. Elsevier; 2013.
Final version of visual abstract.

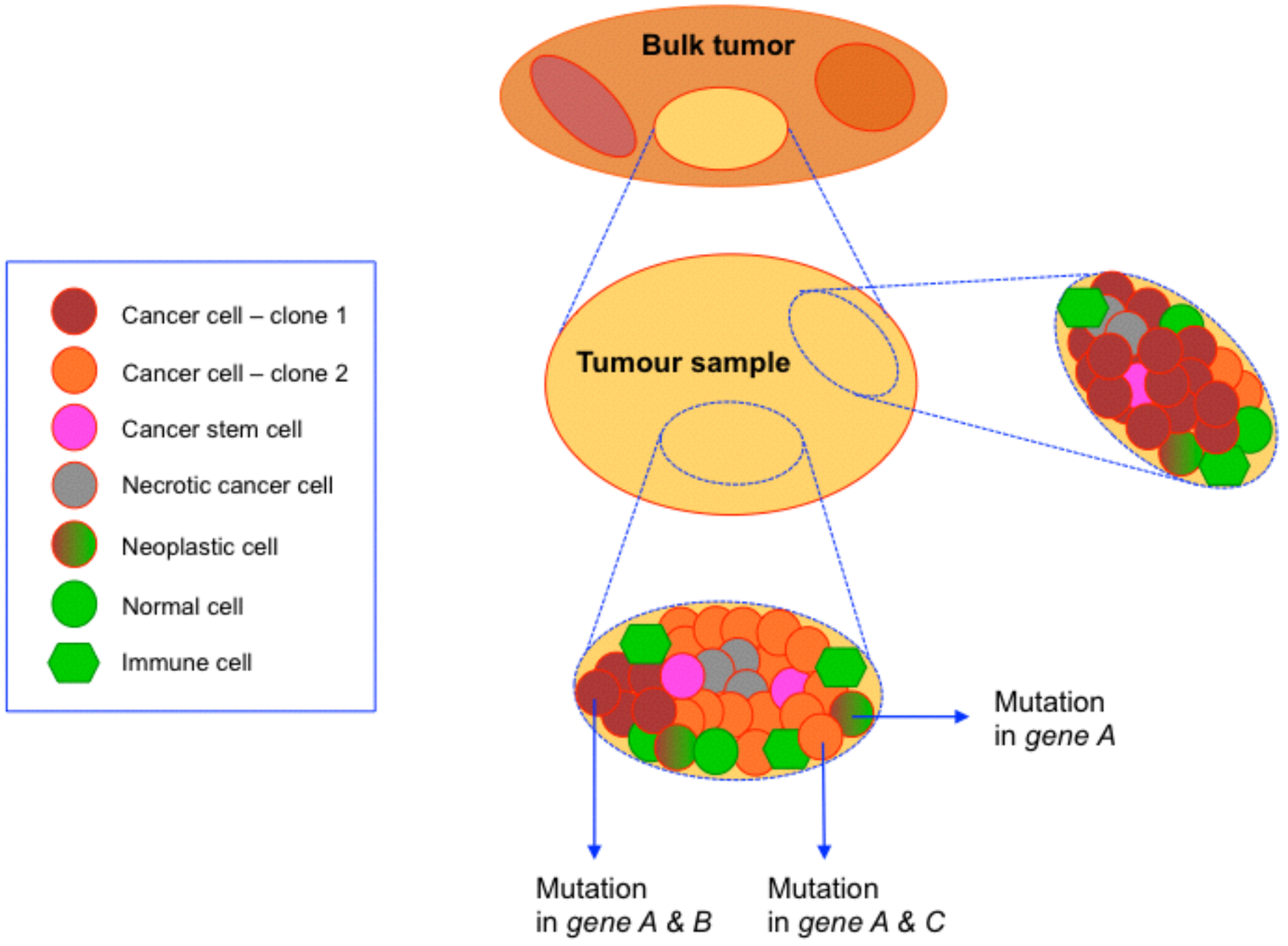




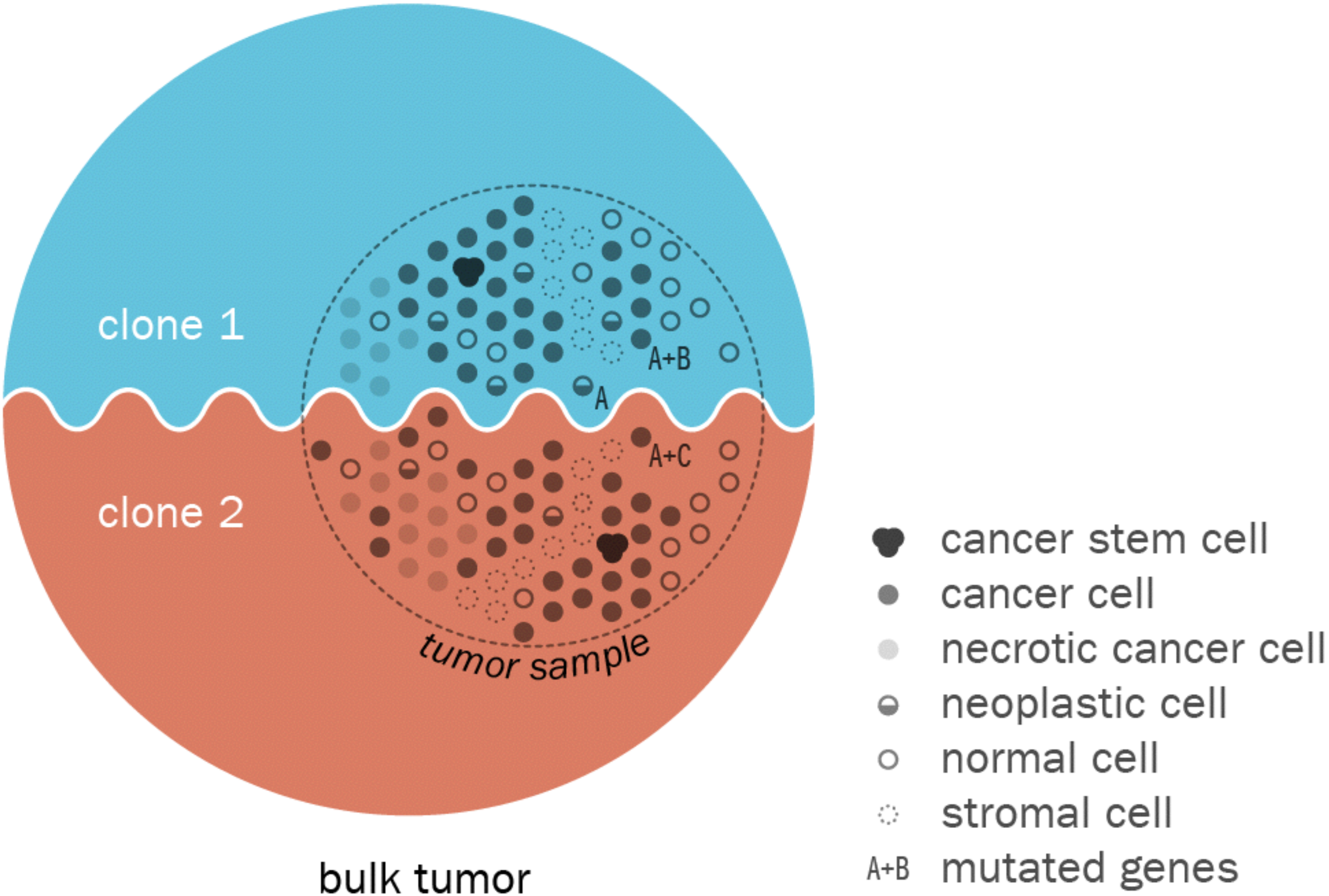
Chun HJE, Khattra J, Krzywinski M, Aparicio SA, Marra MA (Dellaire GD, Berman JN, Arceci RJ. editors). Cancer Genomics. 1st ed. Elsevier; 2013.
Early (top) vs final (bottom) versions of visual abstract.

REDESIGN—IN HOUSE

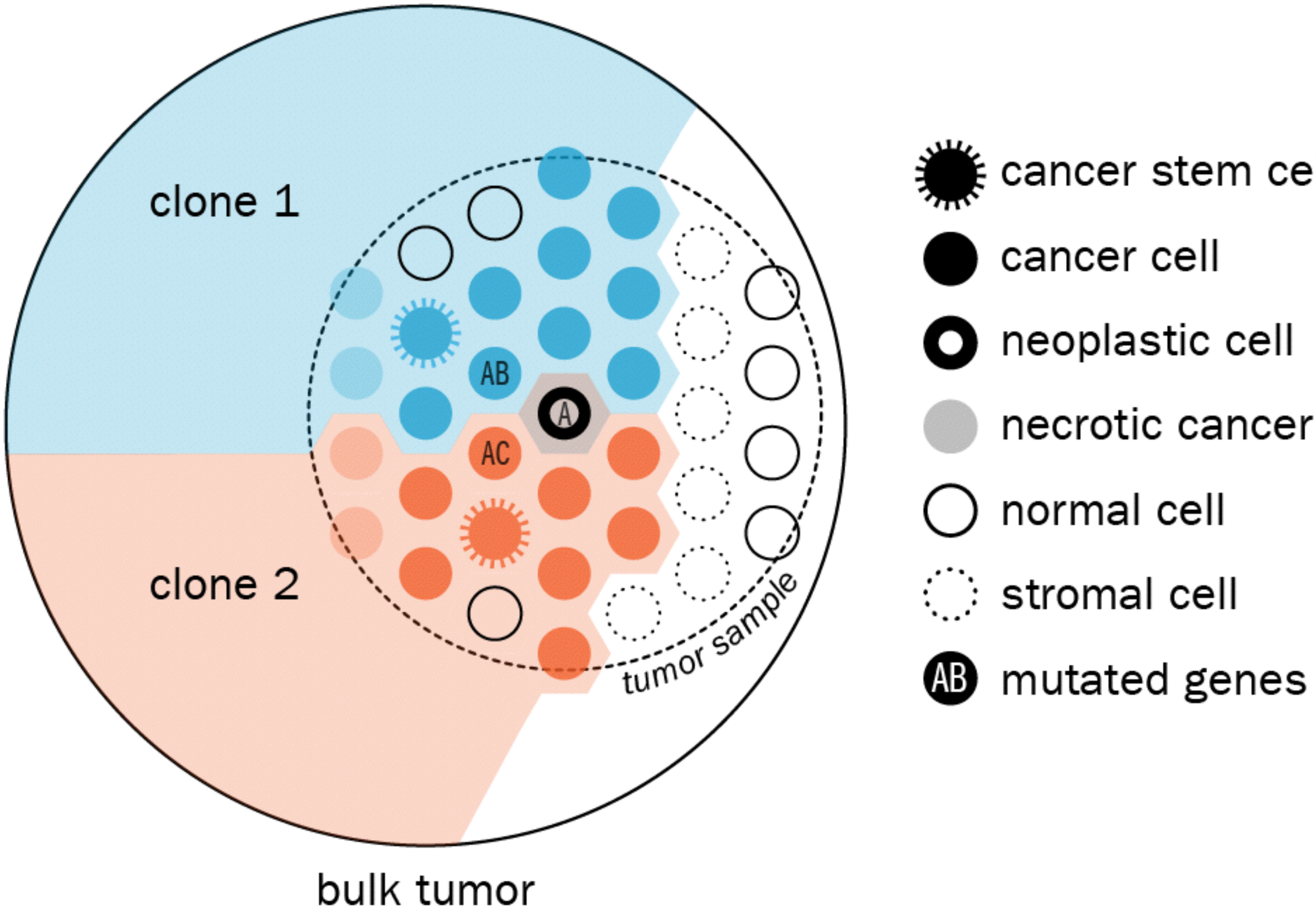
Rational visual vocabulary for cell types.



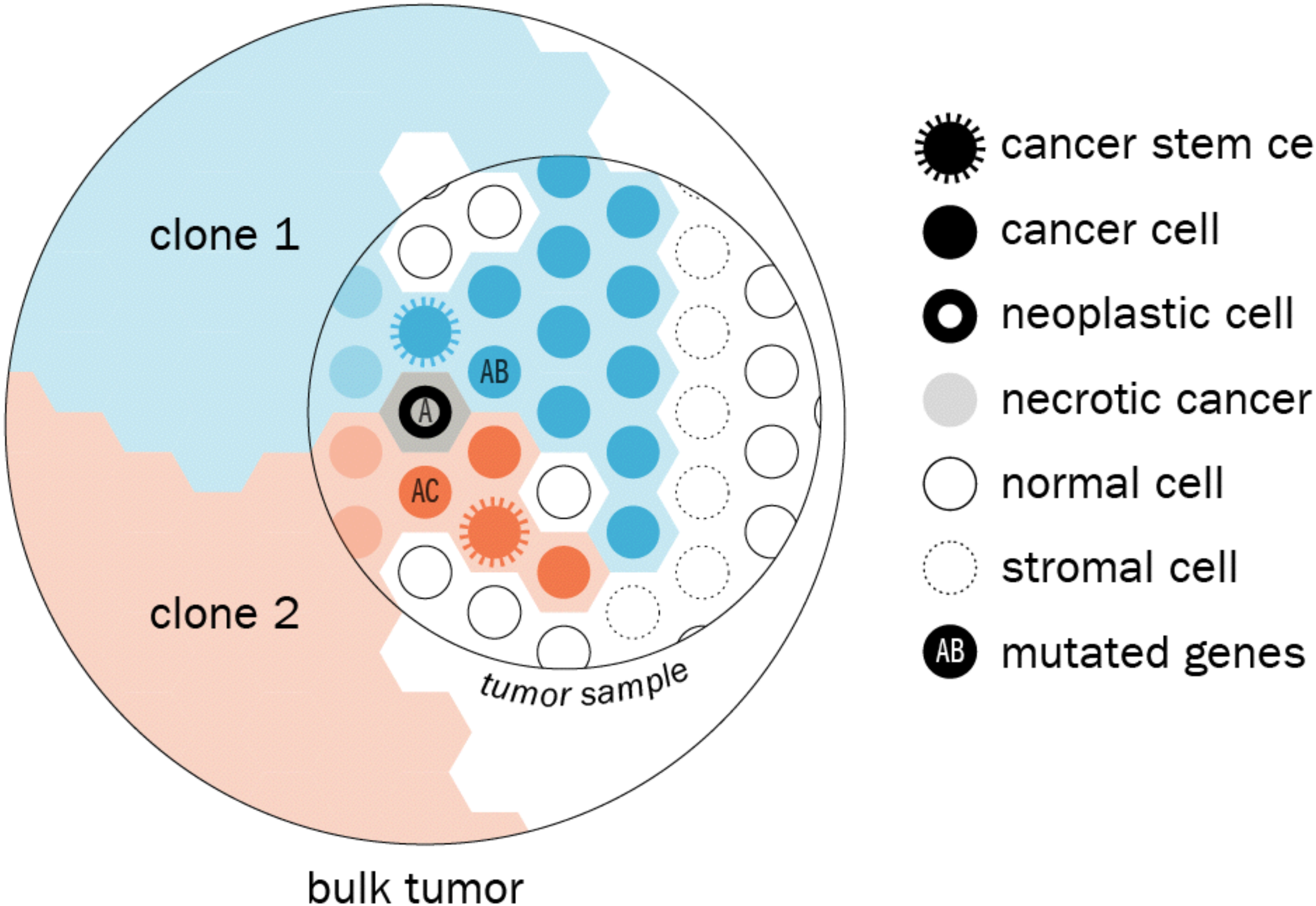
Chun HJE, Khattra J, Krzywinski M, Aparicio SA, Marra MA (Dellaire GD, Berman JN, Arceci RJ. editors). Cancer Genomics. 1st ed. Elsevier; 2013.
Early version of Fig 1.



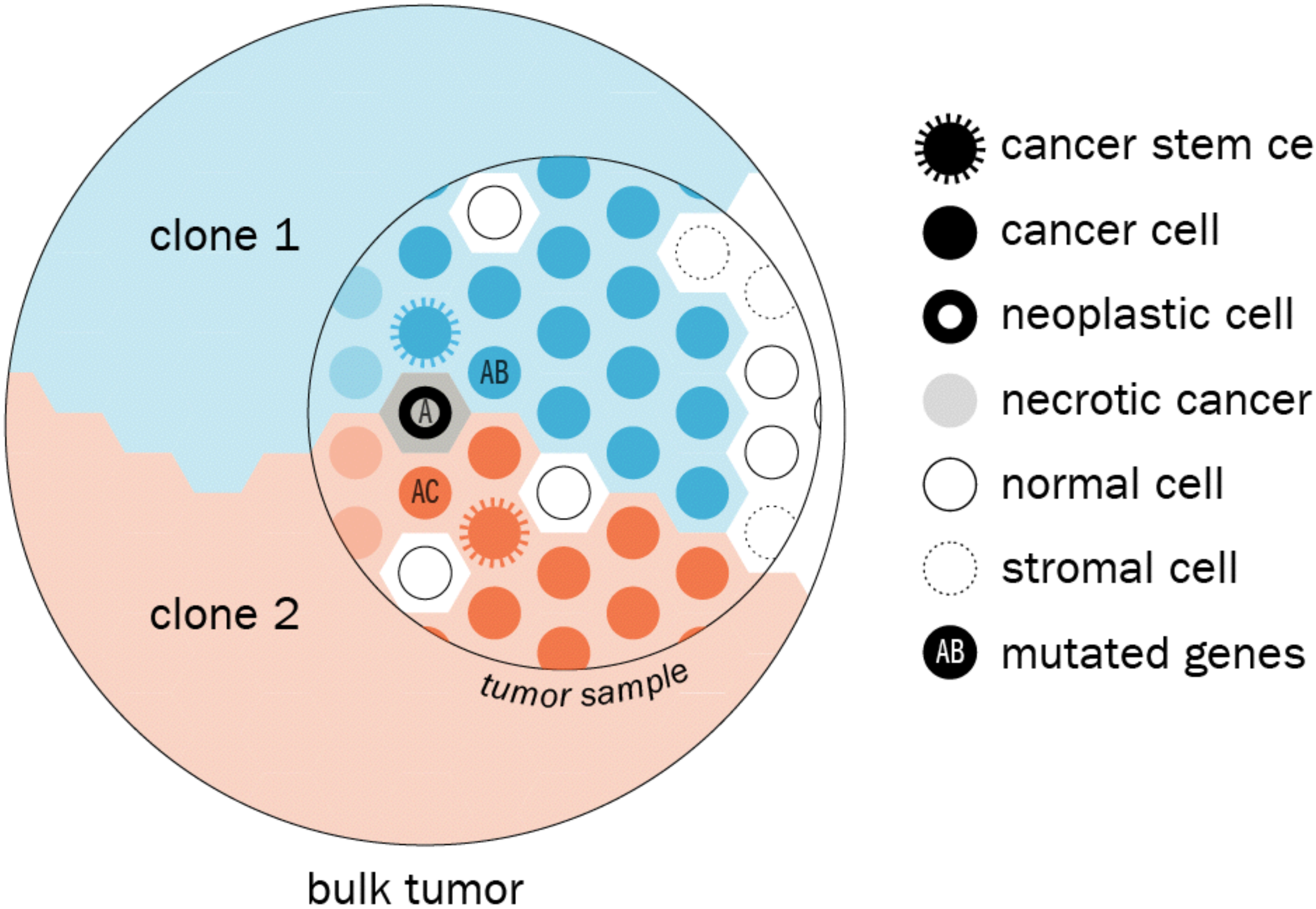
Chun HJE, Khattra J, Krzywinski M, Aparicio SA, Marra MA (Dellaire GD, Berman JN, Arceci RJ. editors). Cancer Genomics. 1st ed. Elsevier; 2013.
Early version of Fig 1.



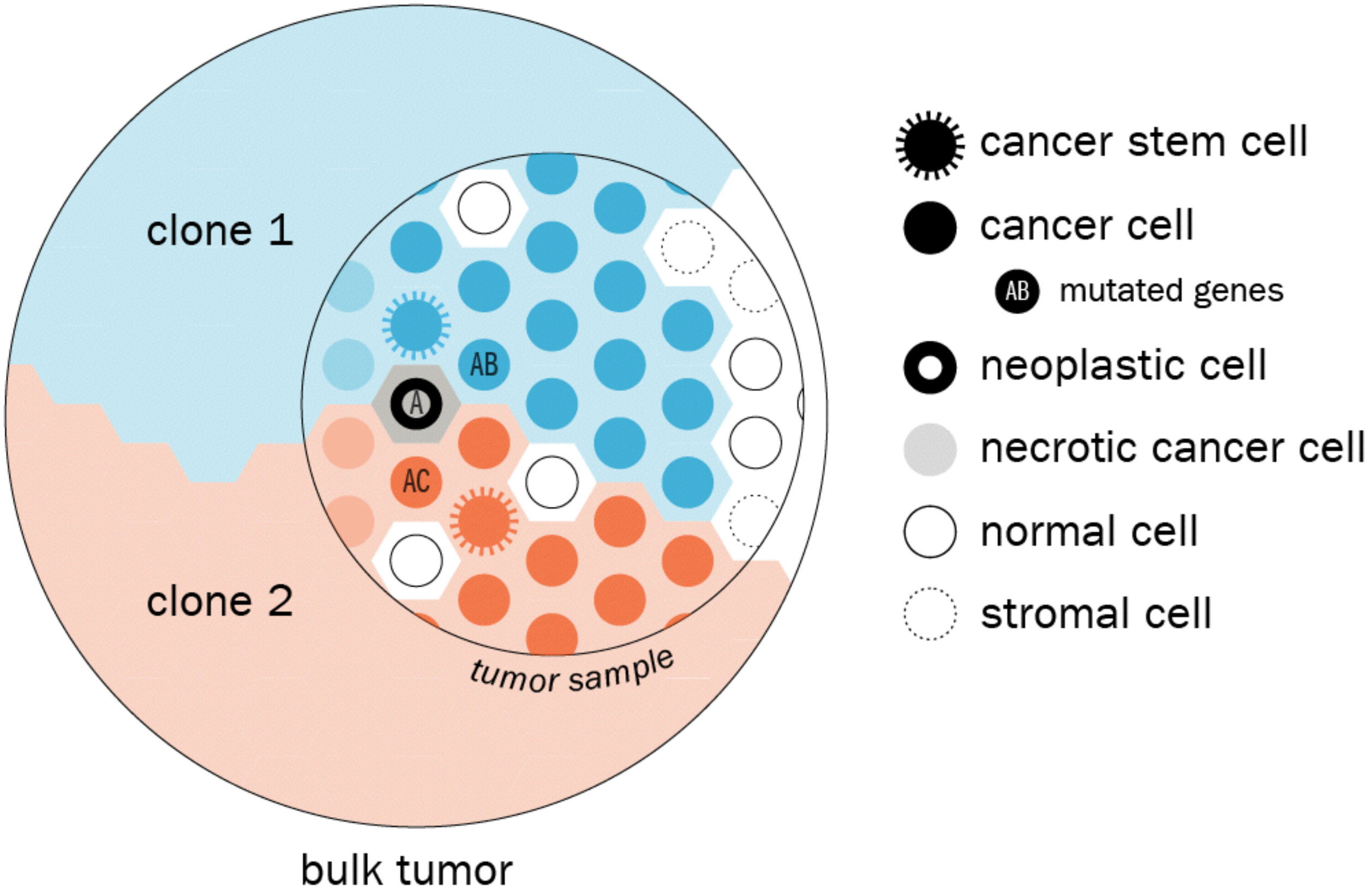
Chun HJE, Khattra J, Krzywinski M, Aparicio SA, Marra MA (Dellaire GD, Berman JN, Arceci RJ. editors). Cancer Genomics. 1st ed. Elsevier; 2013.
Early version of Fig 1.



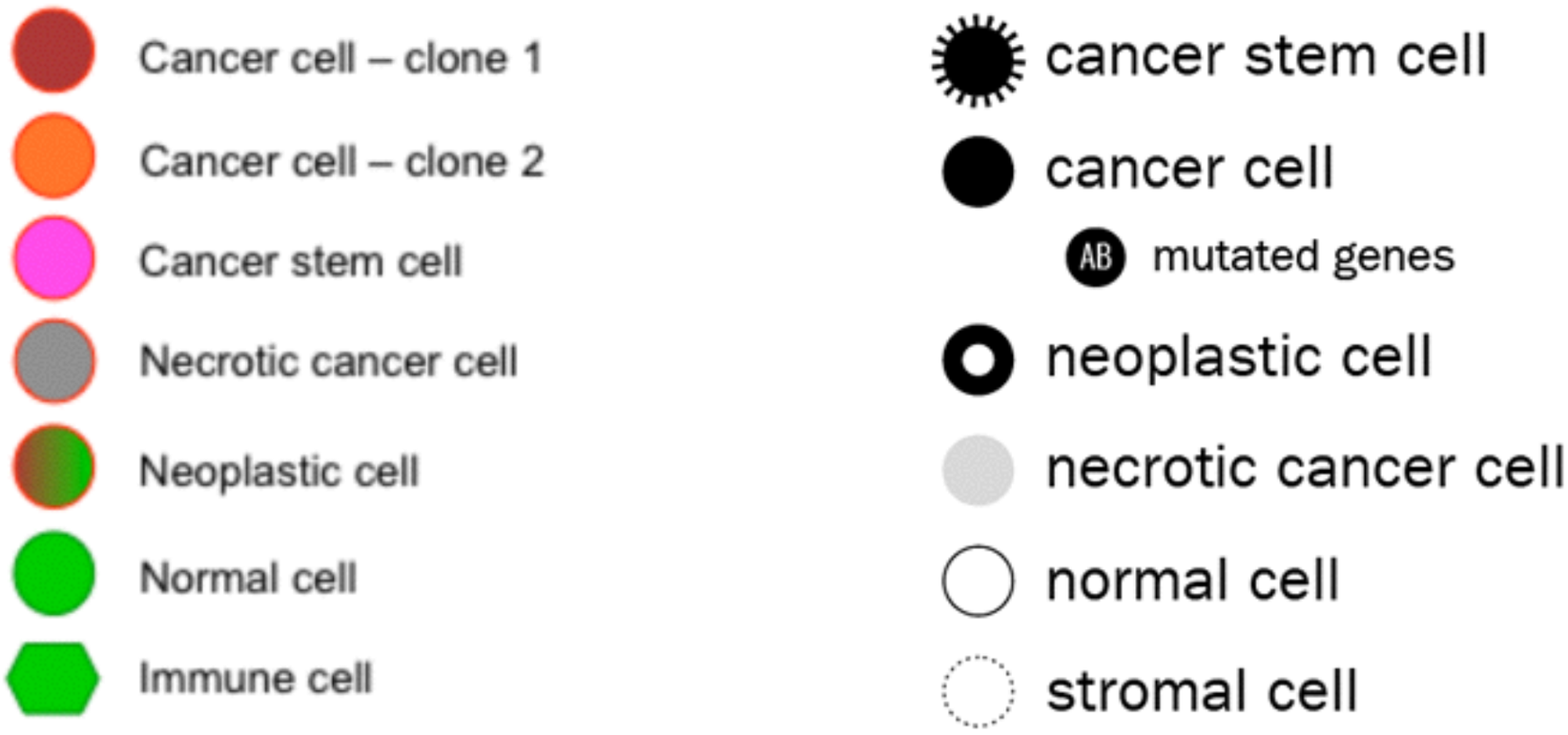
Chun HJE, Khattra J, Krzywinski M, Aparicio SA, Marra MA (Dellaire GD, Berman JN, Arceci RJ. editors). Cancer Genomics. 1st ed. Elsevier; 2013.
Early version of Fig 1.



Chun HJE, Khattra J, Krzywinski M, Aparicio SA, Marra MA (Dellaire GD, Berman JN, Arceci RJ. editors). Cancer Genomics. 1st ed. Elsevier; 2013.
Early version of Fig 1.



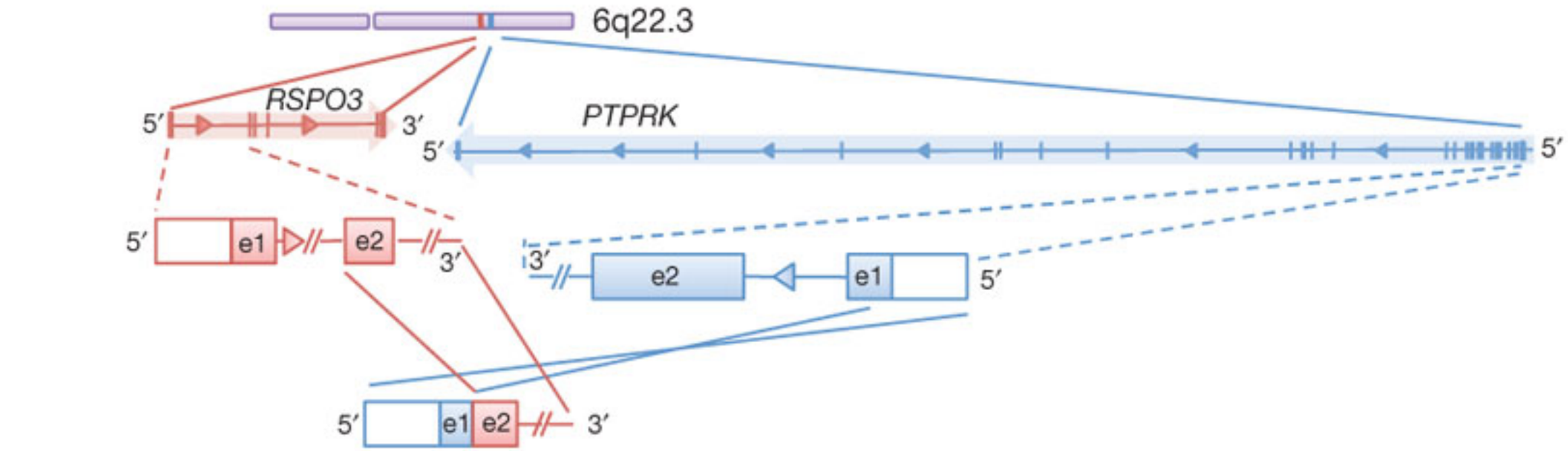
Chun HJE, Khattra J, Krzywinski M, Aparicio SA, Marra MA (Dellaire GD, Berman JN, Arceci RJ. editors). Cancer Genomics. 1st ed. Elsevier; 2013.
Early version of Fig 1.



Chun HJE, Khattra J, Krzywinski M, Aparicio SA, Marra MA (Dellaire GD, Berman JN, Arceci RJ. editors). Cancer Genomics. 1st ed. Elsevier; 2013.
Early (left) vs final (right) version of Fig 1 legend.

REDESIGN—FROM LITERATURE

Assume your audience is smart but impatient.



TCTCCTGGGATCGGCCCAAGGCCAGTTCTCCGCAG|TGCATCCTAACGTTAGTCAAGGCTGCCAAGGAGGCTGTGC

CTCCTGGGATCGGCCCAAGGCCAGTTCTCCGCAG|TGCATCCTAACGTTAGTCAAGGCTGCCAAGGAGGCTGTGCA

TCCTGGGATCGGCCCAAGGCCAGTTCTCCGCAG|TGCATCCTAACGTTAGTCAAGGCTGCCAAGGAGGCTGTGCAA

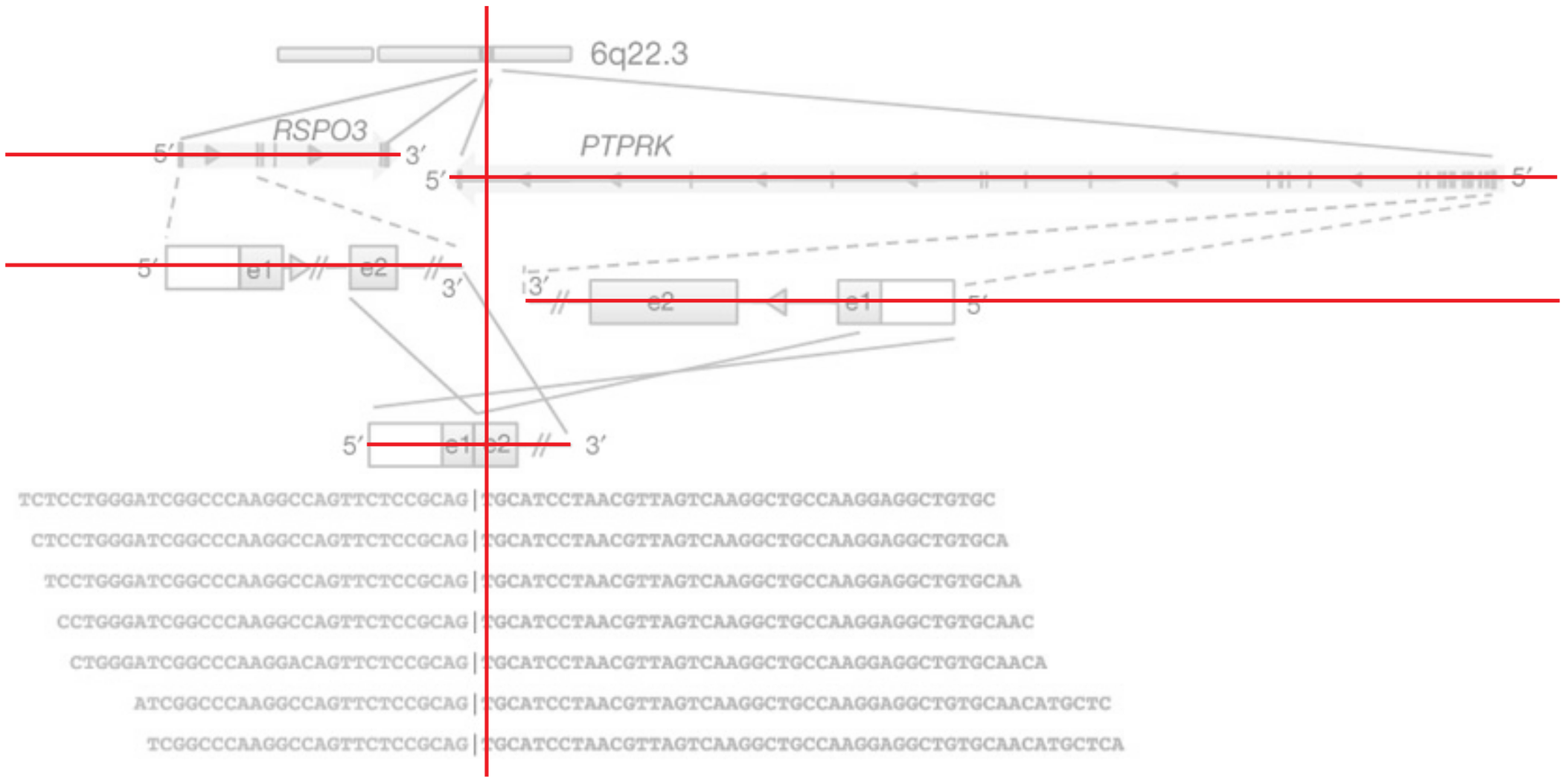
CCTGGGATCGGCCCAAGGCCAGTTCTCCGCAG|TGCATCCTAACGTTAGTCAAGGCTGCCAAGGAGGCTGTGCAAC

CTGGGATCGGCCCAAGGACAGTTCTCCGCAG|TGCATCCTAACGTTAGTCAAGGCTGCCAAGGAGGCTGTGCAACA

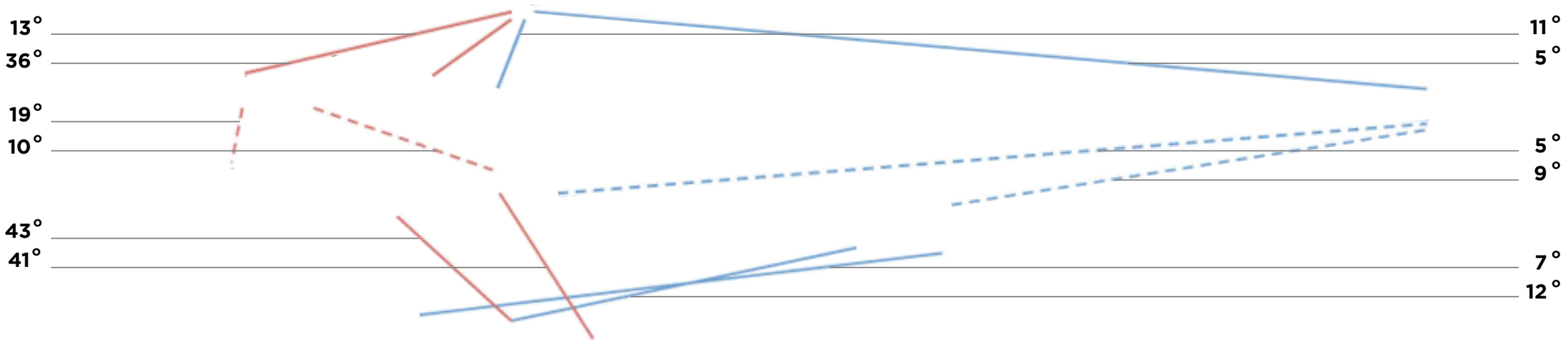
ATCGGCCCAAGGCCAGTTCTCCGCAG|TGCATCCTAACGTTAGTCAAGGCTGCCAAGGAGGCTGTGCAACATGCTC

TCGGCCCAAGGCCAGTTCTCCGCAG|TGCATCCTAACGTTAGTCAAGGCTGCCAAGGAGGCTGTGCAACATGCTCA

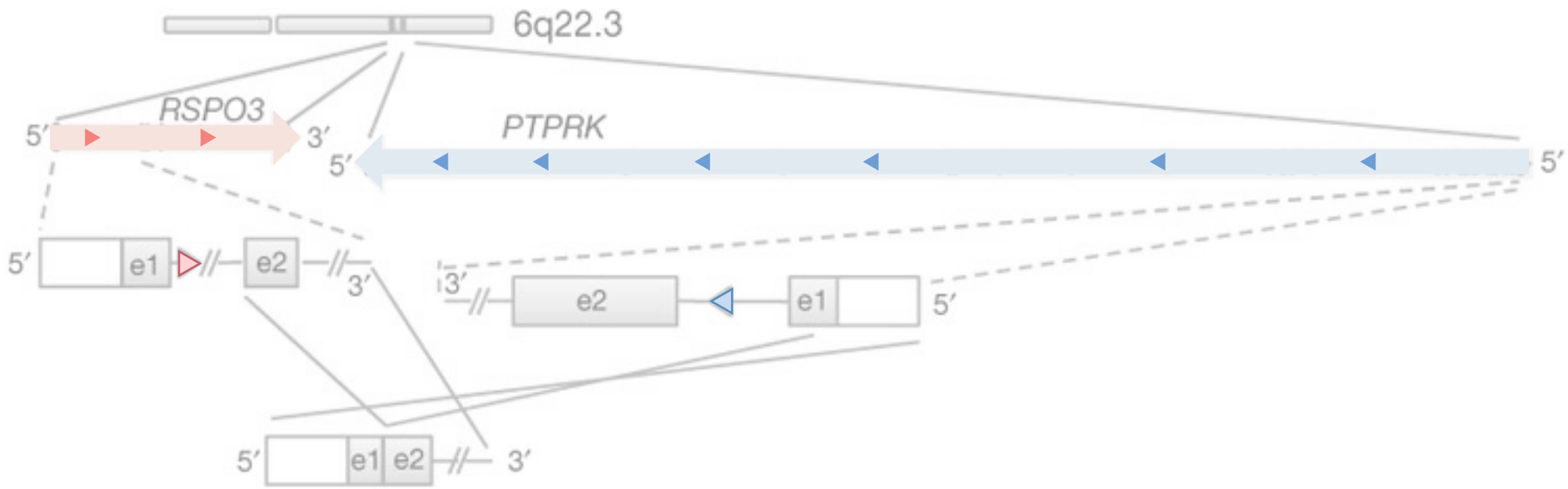
Recurrent R-spondin fusions in colon cancer. Somasekar Seshagiri et al. Nature 488, 660–664 (30 August 2012) doi:10.1038/nature11282



Recurrent R-spondin fusions in colon cancer. Somasekar Seshagiri et al. Nature 488, 660–664 (30 August 2012) doi:10.1038/nature11282



Recurrent R-spondin fusions in colon cancer. Somasekar Seshagiri et al. Nature 488, 660–664 (30 August 2012) doi:10.1038/nature11282



TCTCCTGGGATCGGCCCAAGGCCAGTTCTCCGCAG|TGCATCCTAACGTTAGTCAAGGCTGCCAAGGAGGCTGTGC

CTCCTGGGATCGGCCCAAGGCCAGTTCTCCGCAG|TGCATCCTAACGTTAGTCAAGGCTGCCAAGGAGGCTGTGCA

TCCTGGGATCGGCCCAAGGCCAGTTCTCCGCAG|TGCATCCTAACGTTAGTCAAGGCTGCCAAGGAGGCTGTGCAA

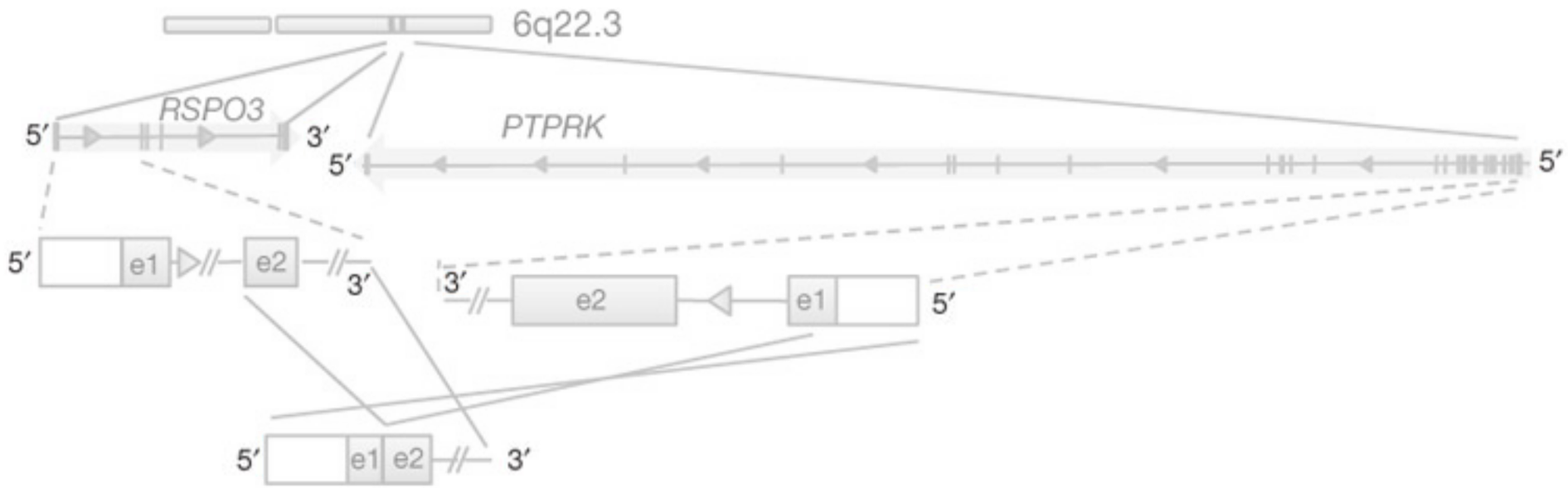
CCTGGGATCGGCCCAAGGCCAGTTCTCCGCAG|TGCATCCTAACGTTAGTCAAGGCTGCCAAGGAGGCTGTGCAAC

CTGGGATCGGCCCAAGGACAGTTCTCCGCAG|TGCATCCTAACGTTAGTCAAGGCTGCCAAGGAGGCTGTGCAACA

ATCGGCCCAAGGCCAGTTCTCCGCAG|TGCATCCTAACGTTAGTCAAGGCTGCCAAGGAGGCTGTGCAACATGCTC

TCGGCCCAAGGCCAGTTCTCCGCAG|TGCATCCTAACGTTAGTCAAGGCTGCCAAGGAGGCTGTGCAACATGCTCA

Recurrent R-spondin fusions in colon cancer. Somasekar Seshagiri et al. Nature 488, 660–664 (30 August 2012) doi:10.1038/nature11282



Recurrent R-spondin fusions in colon cancer. Somasekar Seshagiri et al. Nature 488, 660–664 (30 August 2012) doi:10.1038/nature11282

Recurrent R-spondin fusions in colon cancer. Somasekar Seshagiri et al. Nature 488, 660–664 (30 August 2012) doi:10.1038/nature11282

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)

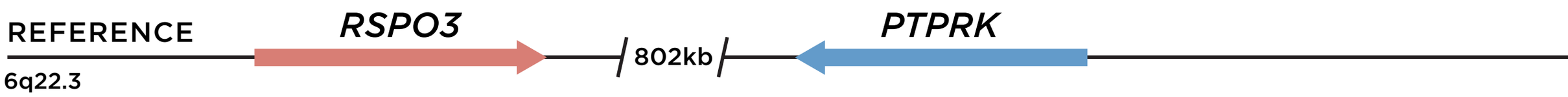
REFERENCE

6q22.3

TCTCCTGGGATCGGCCCAAGGCCAGTTCTCCGCAG TGCATCCTAACGTTAGTCAAGGCTGCCAAGGAGGCTGTGCAACATGCTCA

SAMPLE

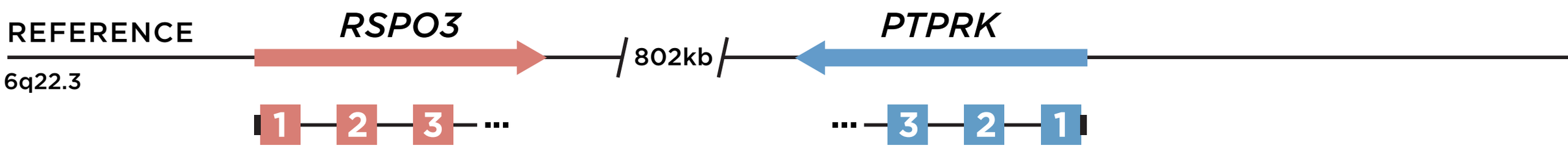
Recurrent R-spondin fusions in colon cancer. Somasekar Seshagiri et al. Nature 488, 660–664 (30 August 2012) doi:10.1038/nature11282
Redesign by M Krzywinski



TCTCCTGGGATCGGCCCAAGGCCAGTTCTCCGCAG TGCATCCTAACGTTAGTCAAGGCTGCCAAGGAGGCTGTGCAACATGCTCA

SAMPLE

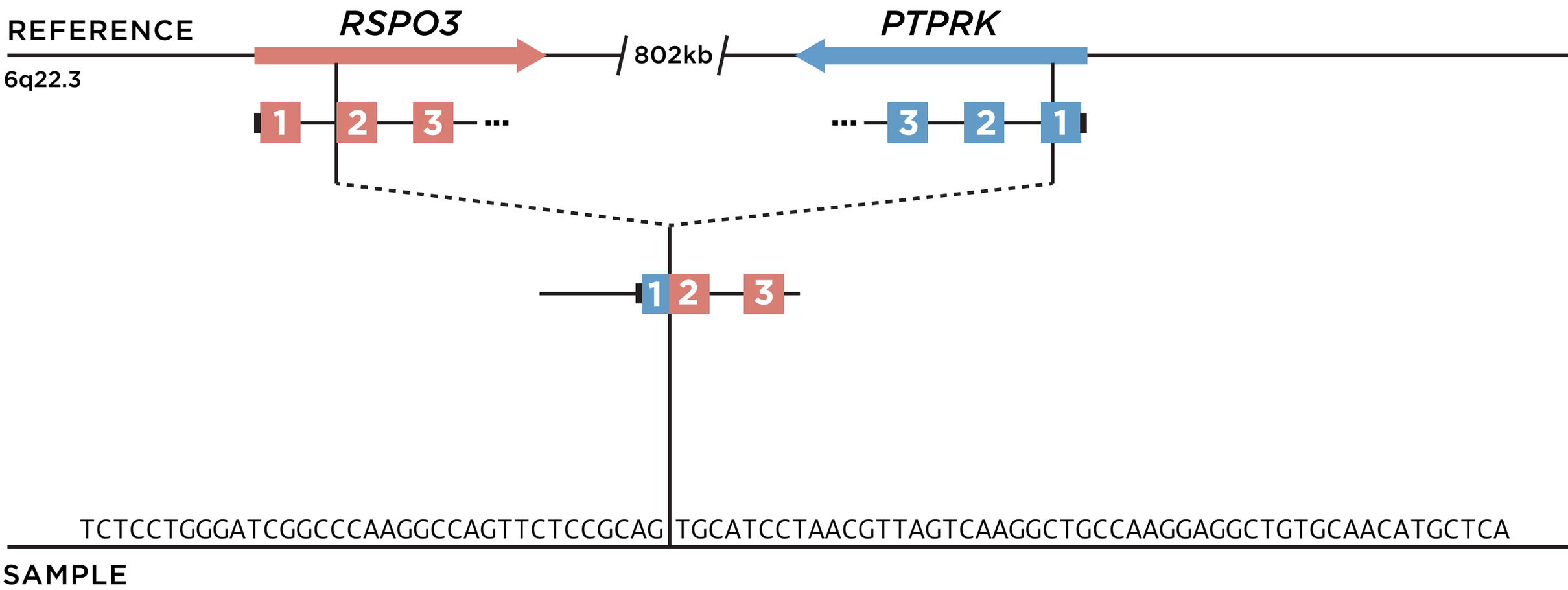
Recurrent R-spondin fusions in colon cancer. Somasekar Seshagiri et al. Nature 488, 660–664 (30 August 2012) doi:10.1038/nature11282
Redesign by M Krzywinski



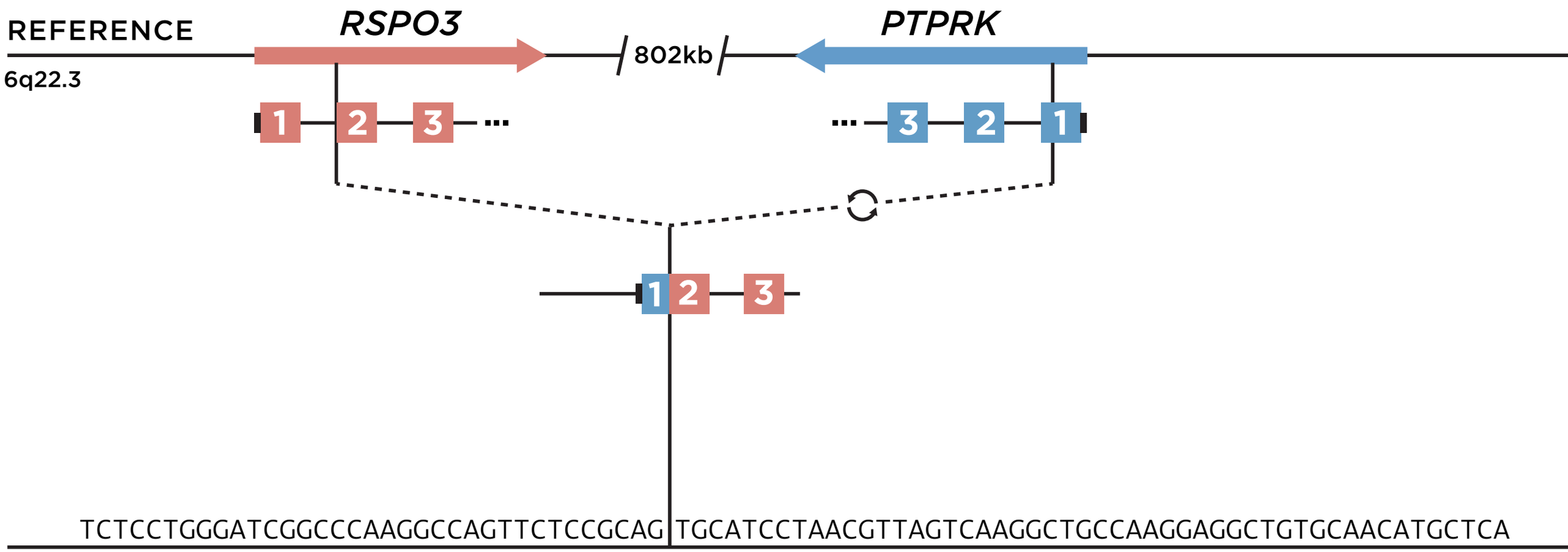
TCTCCTGGGATCGGCCCAAGGCCAGTTCTCCGCAG TGCATCCTAACGTTAGTCAAGGCTGCCAAGGAGGCTGTGCAACATGCTCA

SAMPLE

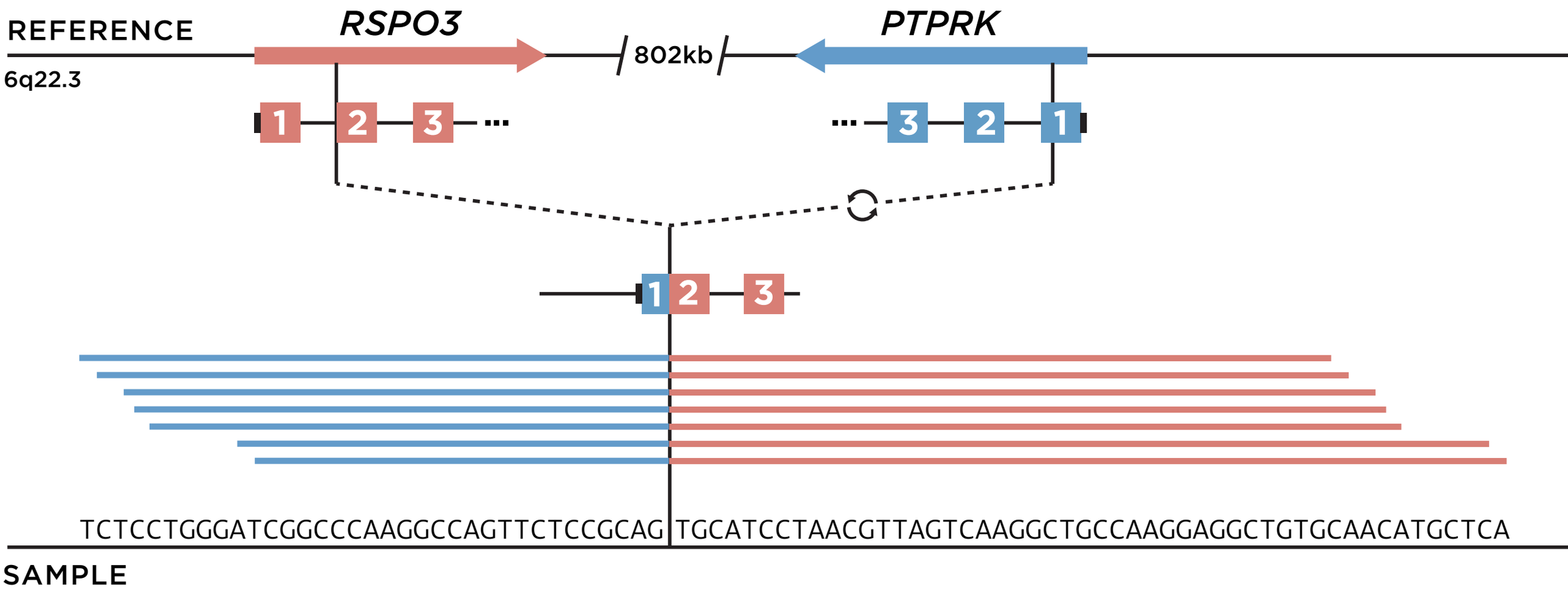
Recurrent R-spondin fusions in colon cancer. Somasekar Seshagiri et al. Nature 488, 660–664 (30 August 2012) doi:10.1038/nature11282
Redesign by M Krzywinski



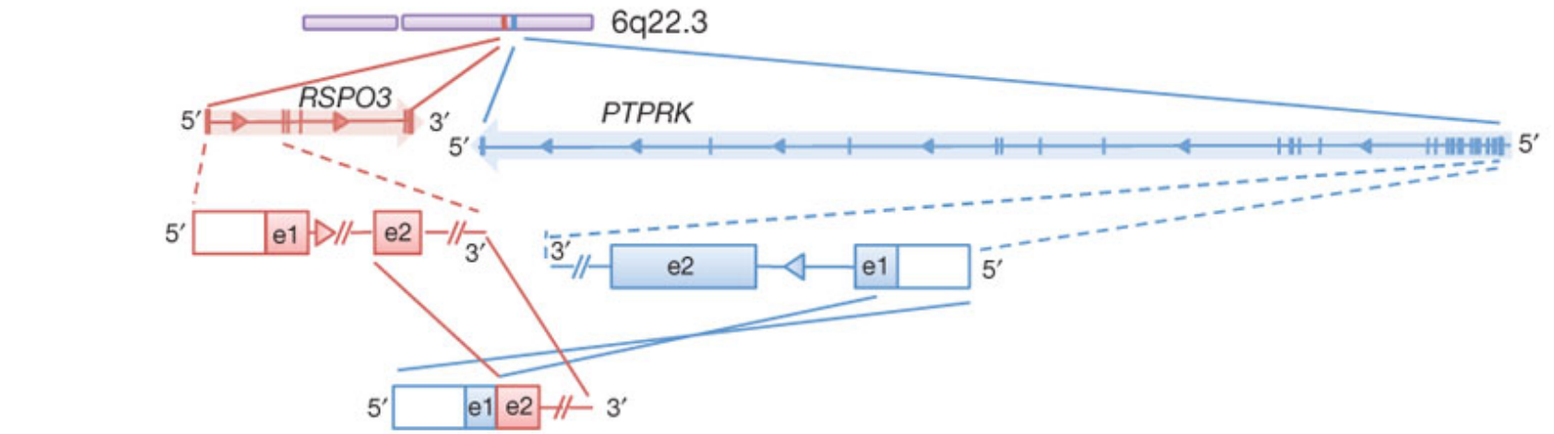
Recurrent R-spondin fusions in colon cancer. Somasekar Seshagiri et al. Nature 488, 660–664 (30 August 2012) doi:10.1038/nature11282
Redesign by M Krzywinski



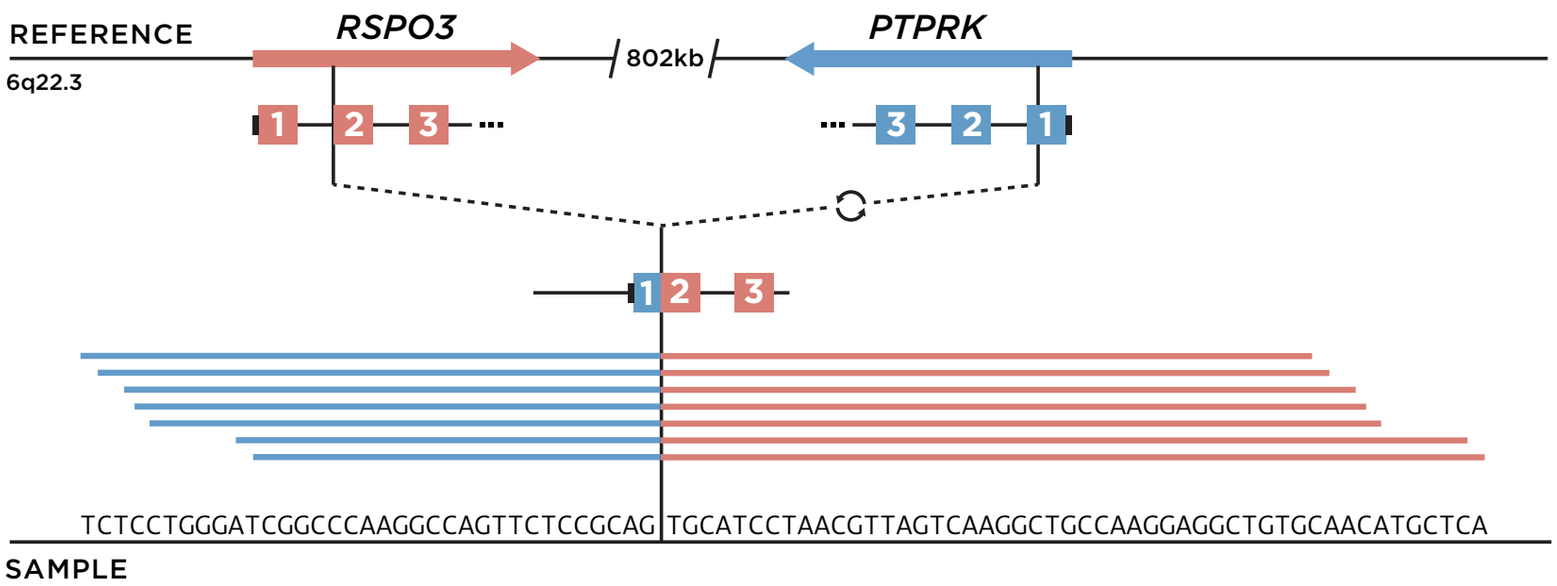
Recurrent R-spondin fusions in colon cancer. Somasekar Seshagiri et al. Nature 488, 660–664 (30 August 2012) doi:10.1038/nature11282
Redesign by M Krzywinski



Recurrent R-spondin fusions in colon cancer. Somasekar Seshagiri et al. Nature 488, 660–664 (30 August 2012) doi:10.1038/nature11282
Redesign by M Krzywinski



TCTCCTGGGATCGGCCCAAGGCCAGTTCTCCGCAG|TGCATCCTAACGTTAGTCAAGGCTGCCAAGGAGGCTGTGC
CTCCTGGGATCGGCCCAAGGCCAGTTCTCCGCAG|TGCATCCTAACGTTAGTCAAGGCTGCCAAGGAGGCTGTGCA
TCCTGGGATCGGCCCAAGGCCAGTTCTCCGCAG|TGCATCCTAACGTTAGTCAAGGCTGCCAAGGAGGCTGTGCAA
CCTGGGATCGGCCCAAGGCCAGTTCTCCGCAG|TGCATCCTAACGTTAGTCAAGGCTGCCAAGGAGGCTGTGCAAC
CTGGGATCGGCCCAAGGACAGTTCTCCGCAG|TGCATCCTAACGTTAGTCAAGGCTGCCAAGGAGGCTGTGCAACA
ATCGGCCCAAGGCCAGTTCTCCGCAG|TGCATCCTAACGTTAGTCAAGGCTGCCAAGGAGGCTGTGCAACATGCTC
TCGGCCCAAGGCCAGTTCTCCGCAG|TGCATCCTAACGTTAGTCAAGGCTGCCAAGGAGGCTGTGCAACATGCTCA



Recurrent R-spondin fusions in colon cancer. Somasekar Seshagiri et al. Nature 488, 660–664 (30 August 2012) doi:10.1038/nature11282
Redesign by M Krzywinski

Further reading

- VIZBI conference videos and tutorials: http://www.vizbi.org
- Visualization principles tutorial by Jessie Kennedy, Cydney Nielsen and Martin Krzywinski: http://mkweb.bcgsc.ca/vizbi/2012/
- Points of View Column in Nature Methods by Bang Wong: http://bang.clearscience.info/?p=546

Exam-like questions

- Mention at least 3 points that you should focus on when you visualise your data and mention at least 3 points that you should avoid in visualization. Also provide an example to your points.
- What is the problem with the colours? How can you tackle it?
- Why one should not use 3D figures?
- Exercise: you need to "refactor" a figure