THE UNIVERSITY OF
SYDNEY

CANADA'S MICHAEL SMITH
GENOME
SCIENCES
CENTRE

ESSENTIALS OF DATA VISUALIZATION

THINKING ABOUT DRAWING DATA + COMMUNICATING SCIENCE

SHAPES

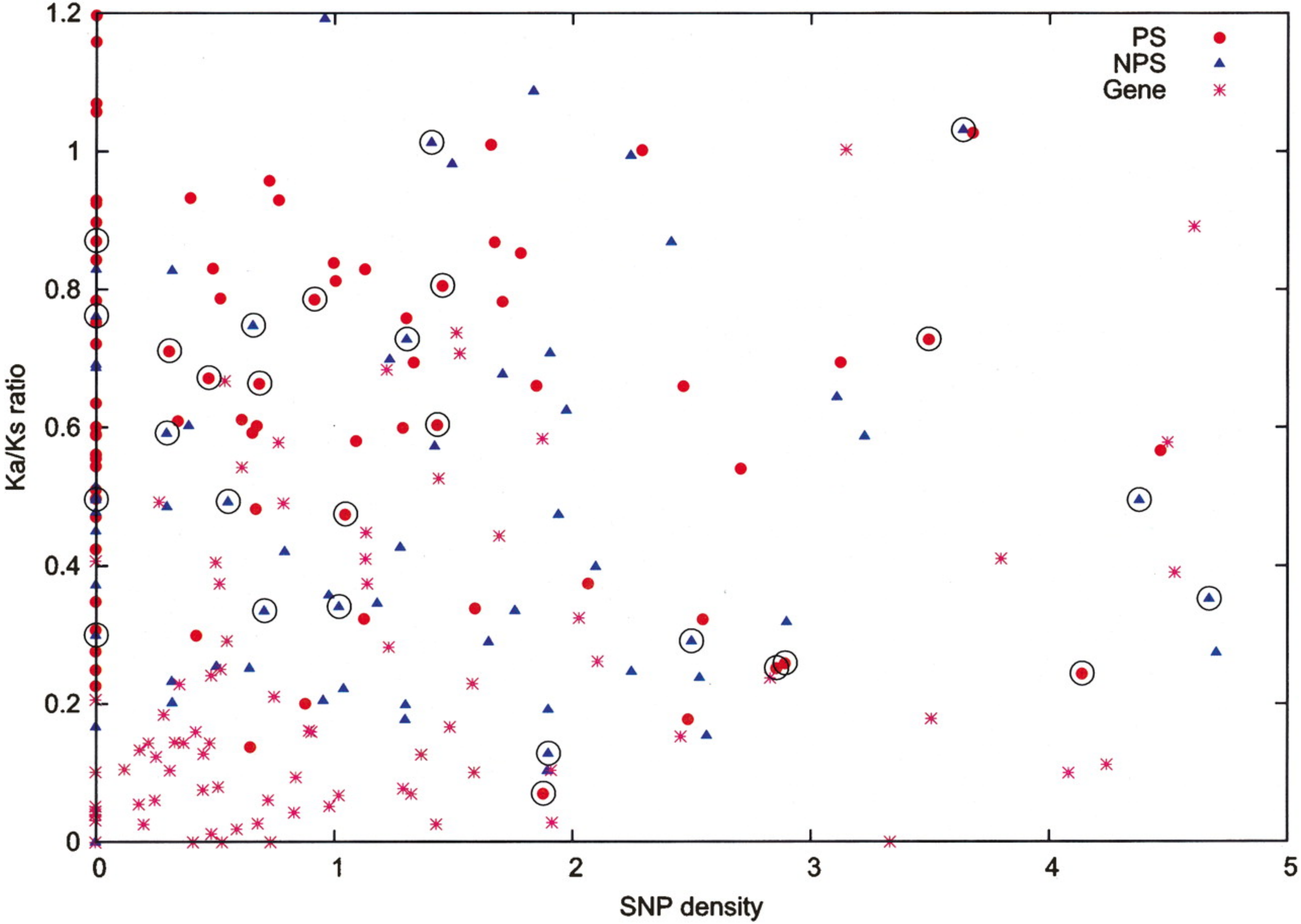
choosing symbols

In the section about encoding, I used circles for the symbols in the scatter plot.

Could I have used squares? Sure. Would it matter in that example? Probably not.

Now, what about triangles, or stars or little flowers? Unicorns?

Maybe you can see where this is going. The simplest shapes are preferable to complex ones. They have fewer complex internal parts so the eye doesn’t get hung up on things like corners and sharp points. Another benefit of the circle is that the intersection between two circles is never a circle—not the case for a square. It’s easy to make a square by intersecting squares, or triangles by intersecting triangles. This is an important point when drawing a lot of data points where there is occlusion.

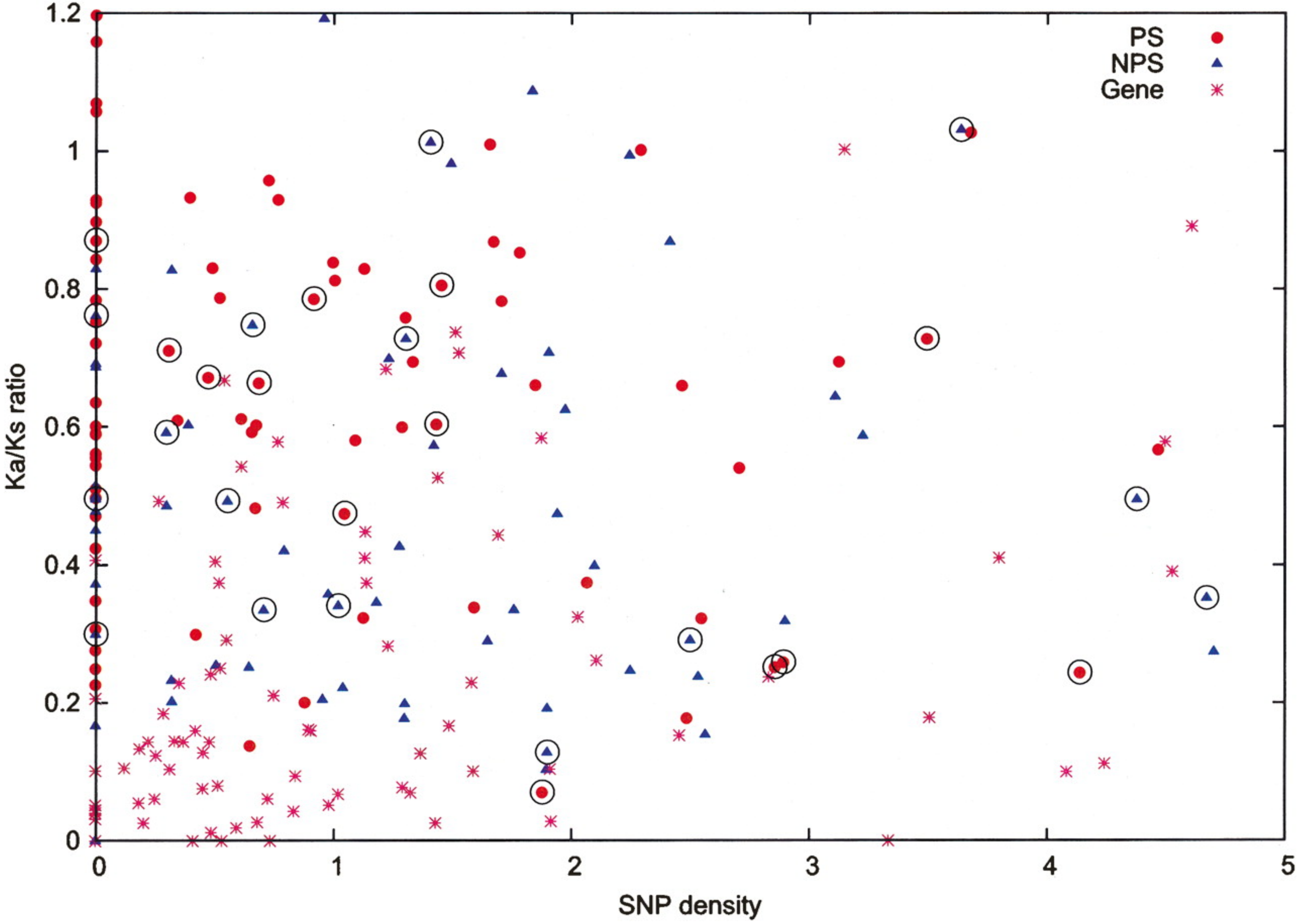


PS ●

NPS ▲

Gene ✳

Genome Res (2007) 17:839–51.



PS ●

NPS ▲

Gene ✳

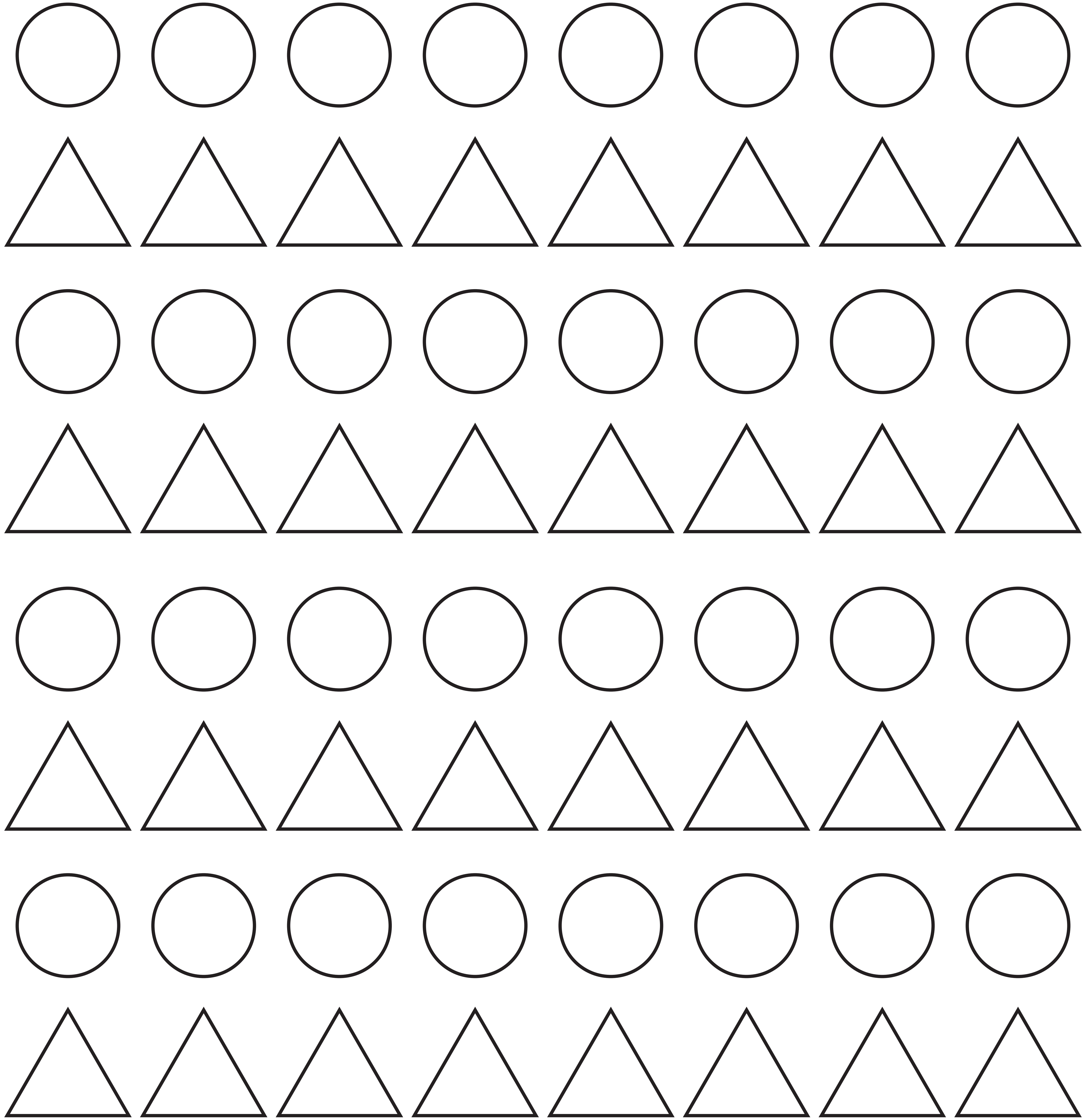
Genome Res (2007) 17:839–51.

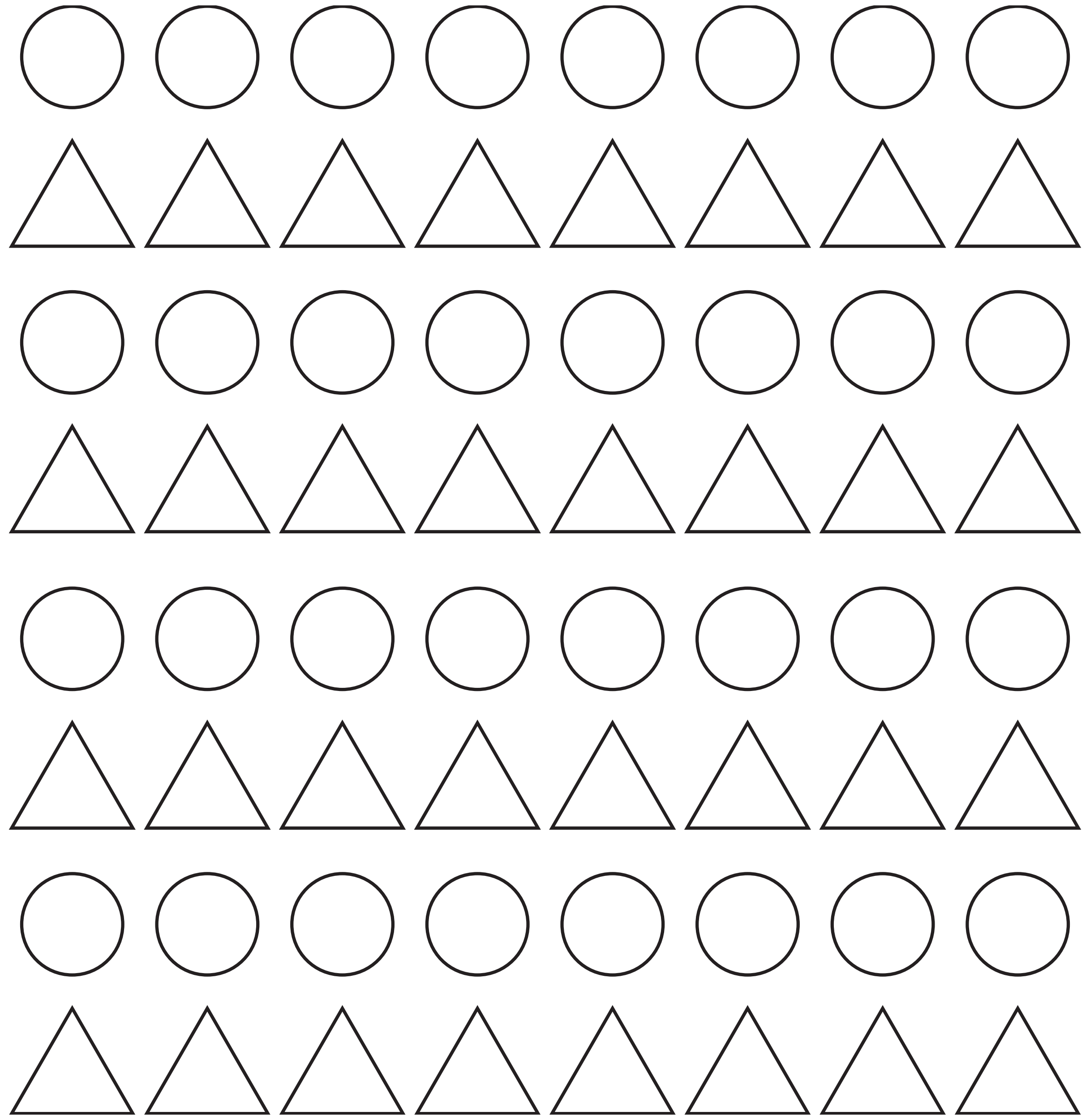
Vision is obviously complicated. The question “how do we see things or information” can’t be answered in a simple way.

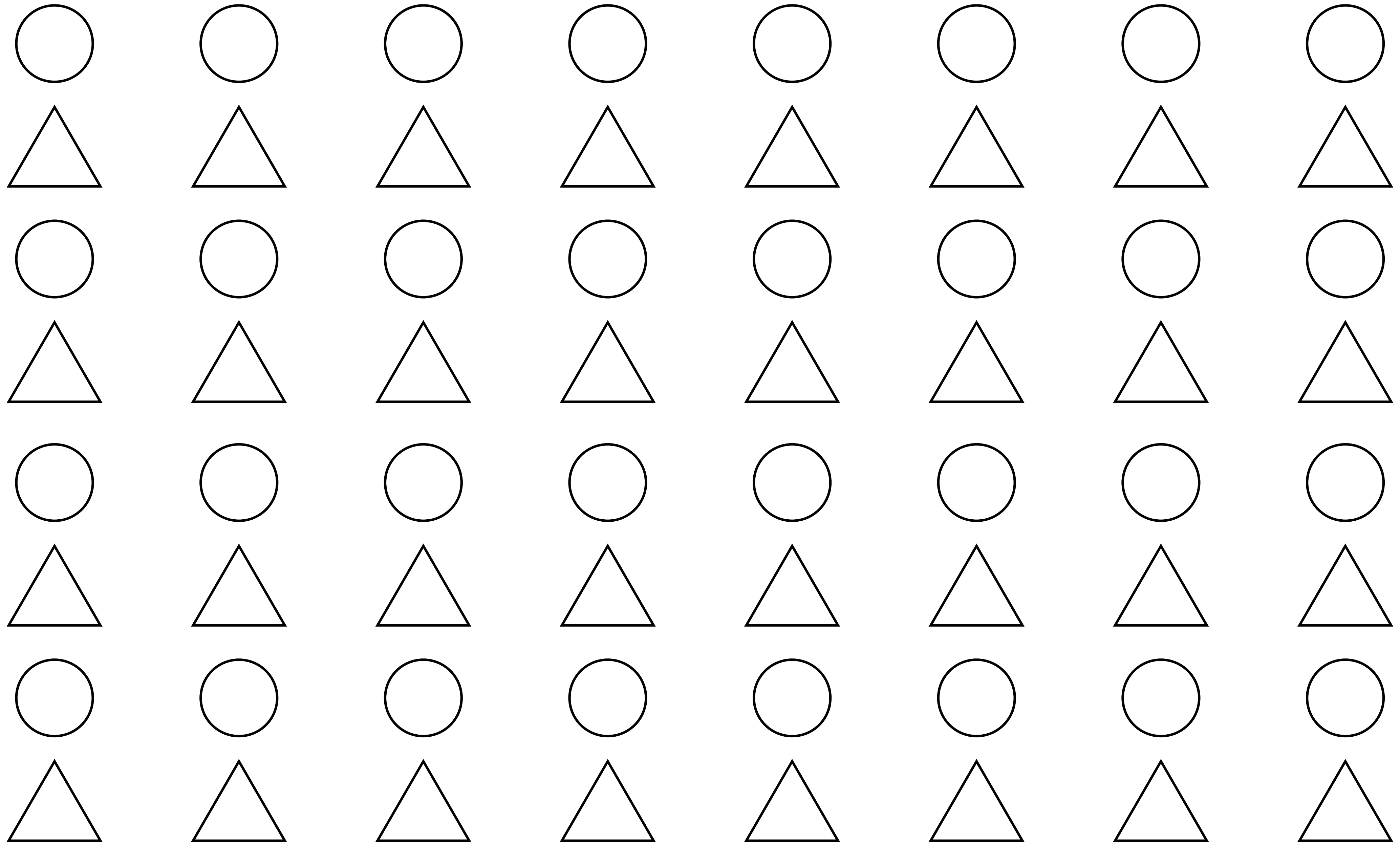
However, there are some broad notions—called Gestalt principles—that describe certain phenomena that are always at play during visual perception.

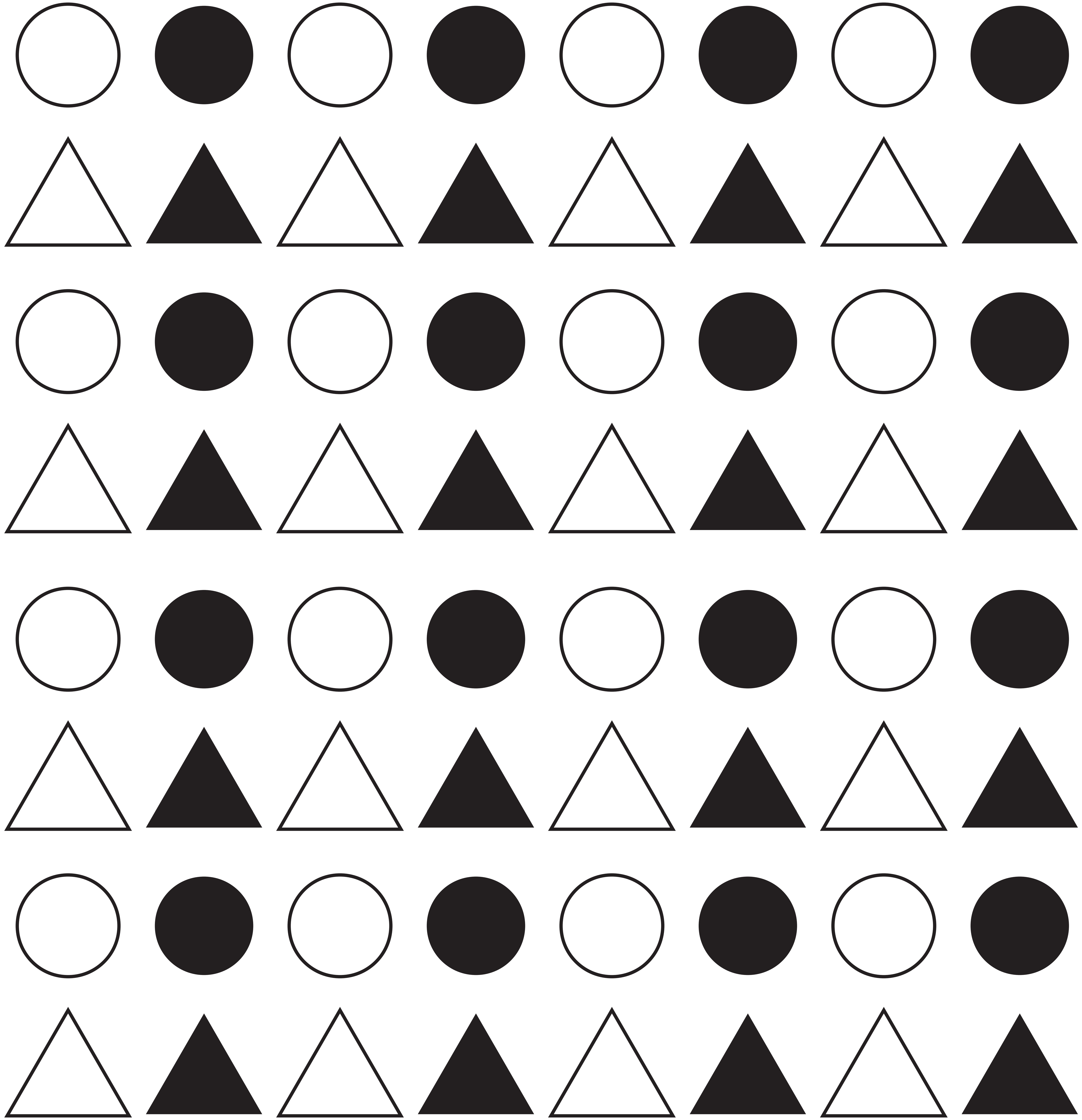
These principles are reasonable and intuitive—you’ve experience them every time you open your eyes. However, they do help us talk about phenomena that are at play.

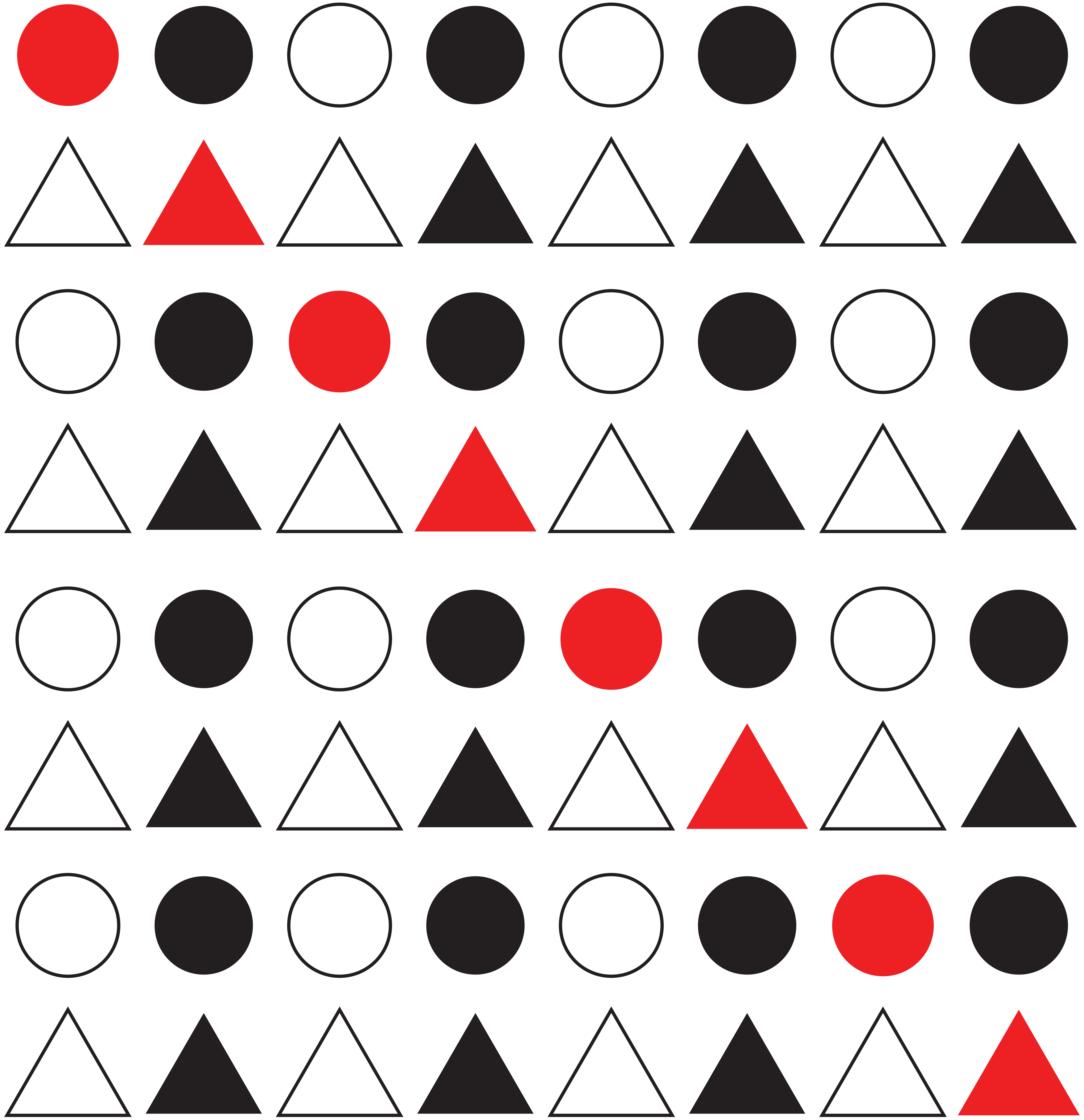
Let’s look at how the principle of similarity and grouping works.

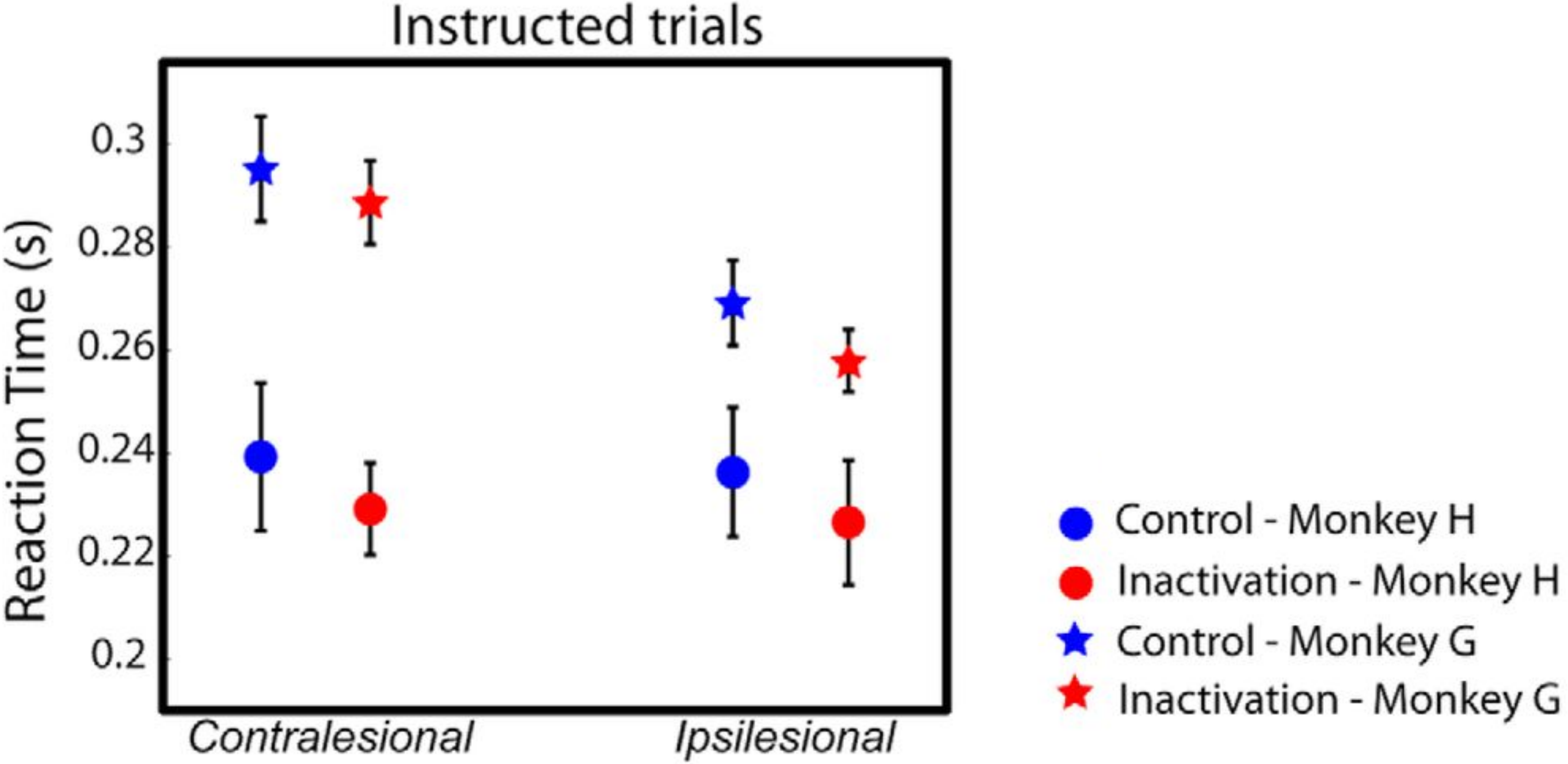












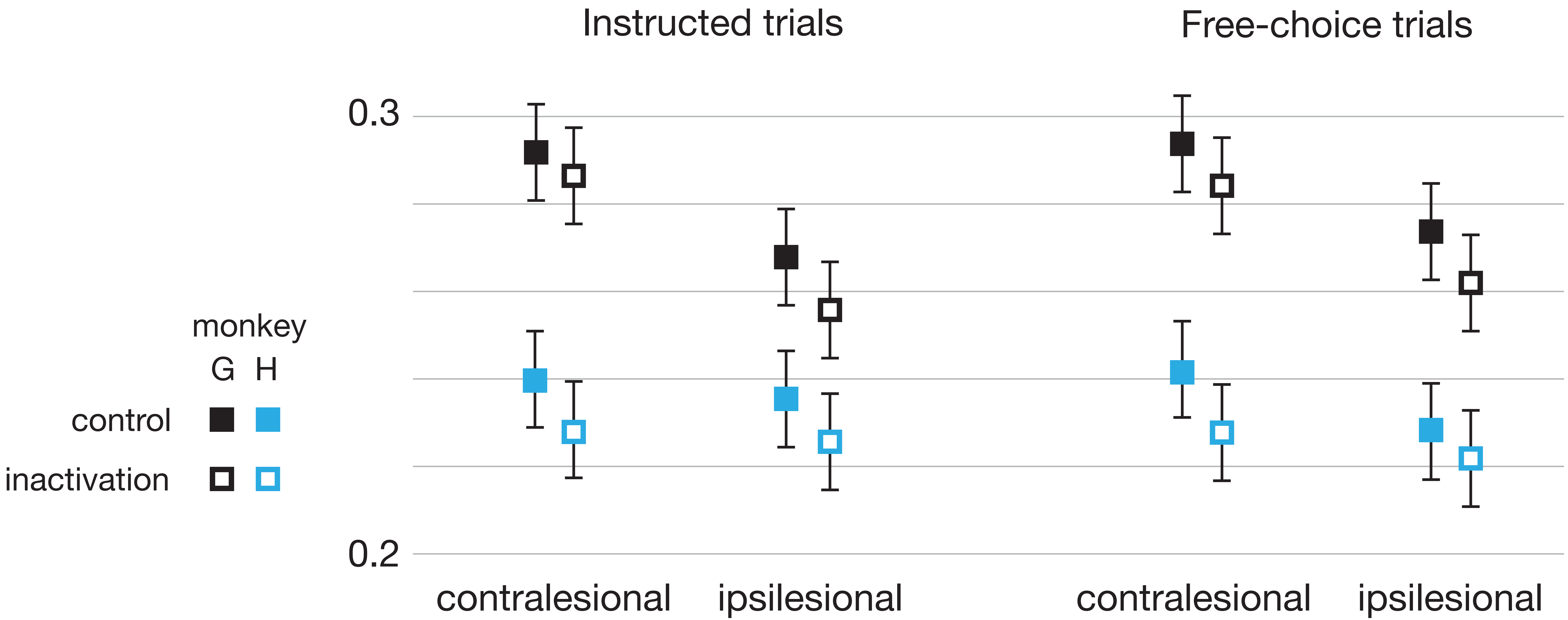
Instructed trials

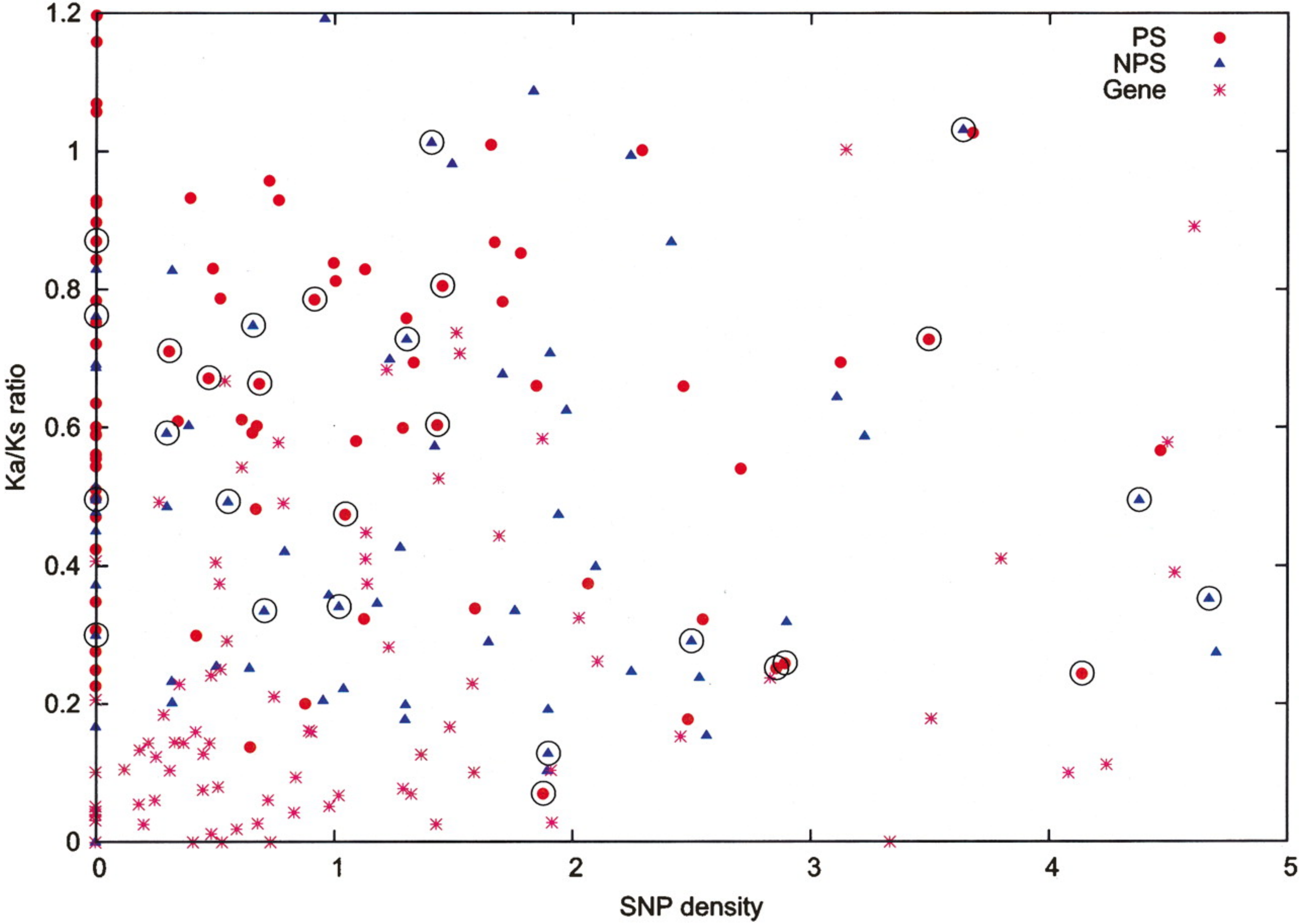
J Neurosci (2015) 35:11719–11728.

- 🔵 control, monkey H
- 🔴 inactivation, monkey H
- ★ control, monkey G
- ★ inactivation, monkey G

	monkey	
	G	H
control	■	🟦
inactivation	□	□

J Neurosci (2015) 35:11719–11728.


Instructed trials
Free-choice trials
0.3
0.2
monkey
G H
control
inactivation
contralesional
ipsilesional
contralesional
ipsilesional



PS ●

NPS ▲

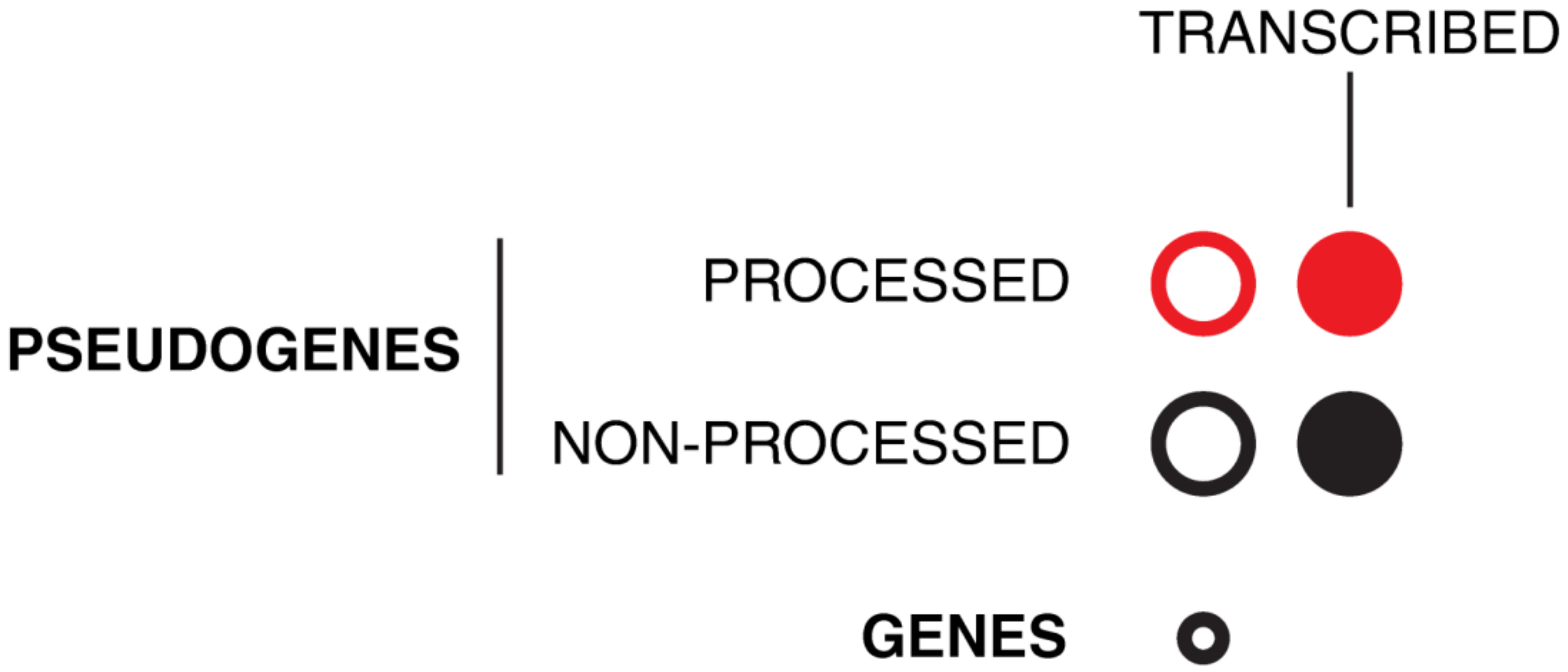
Gene ✳

Genome Res (2007) 17:839–51.

PS ●

NPS ▲

Gene ✳



TRANSCRIBED

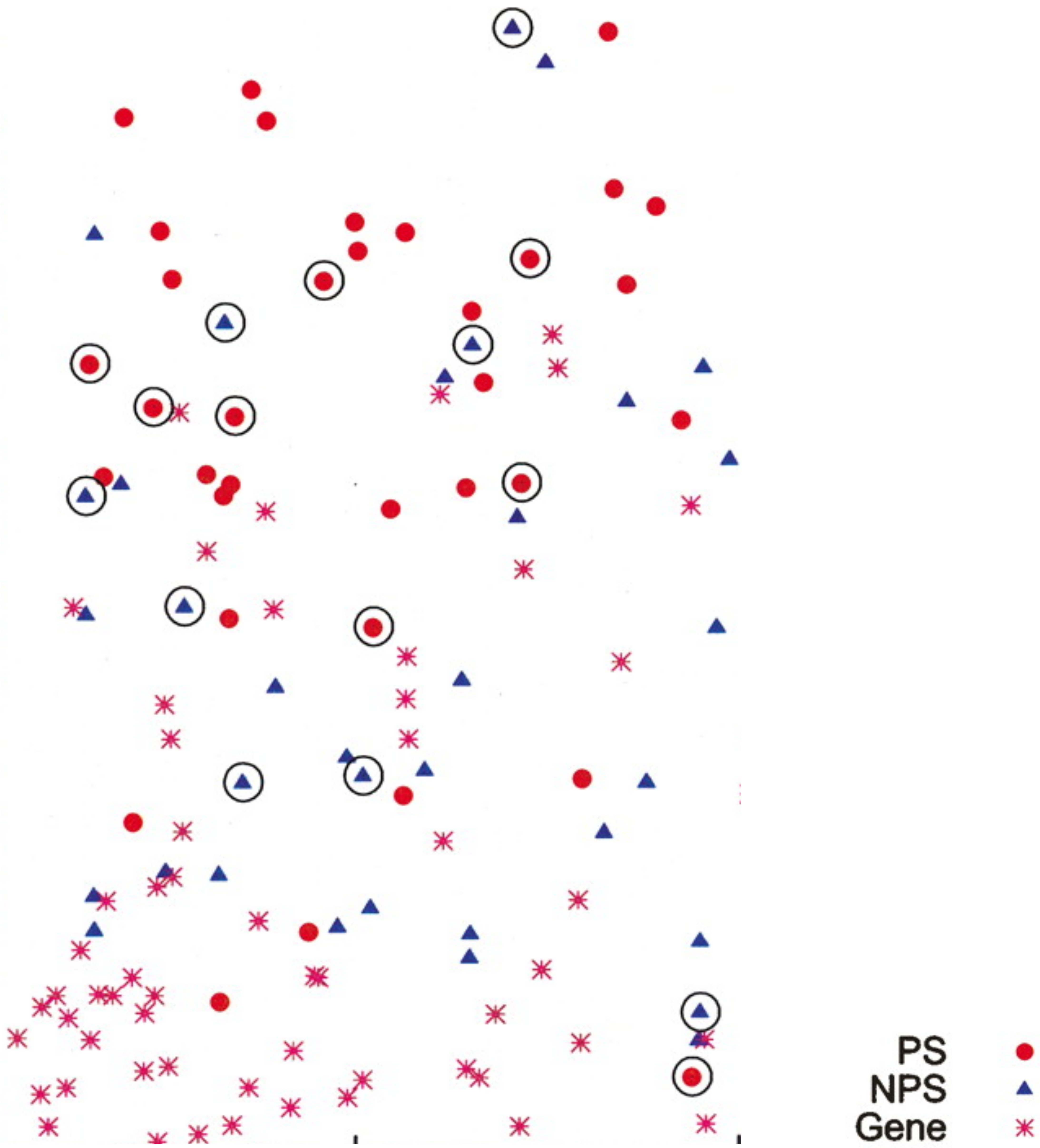
PSEUDOGENES

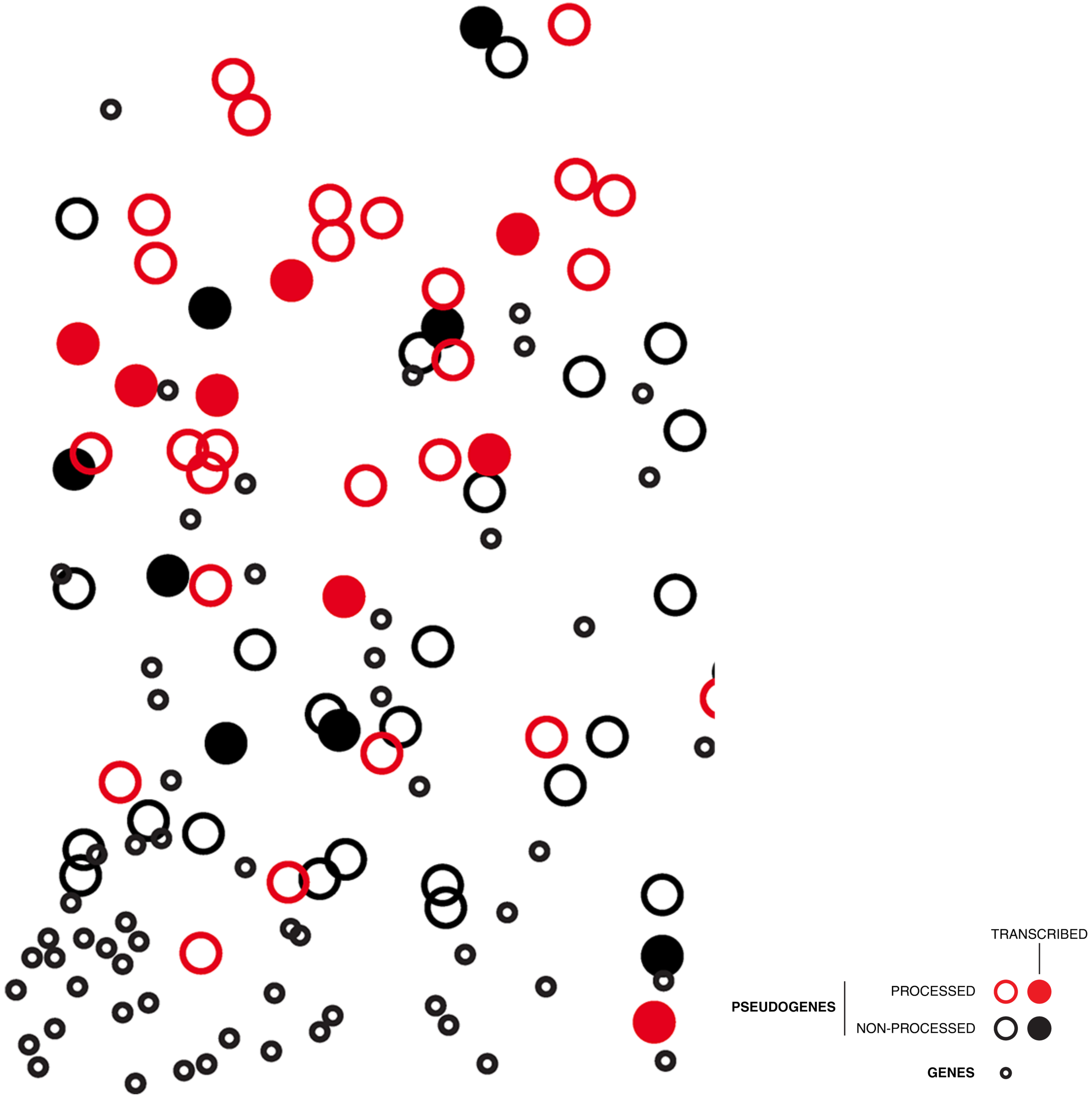
PROCESSED

NON-PROCESSED

GENES

Genome Res (2007) 17:839–51.





Genome Res (2007) 17:839–51.

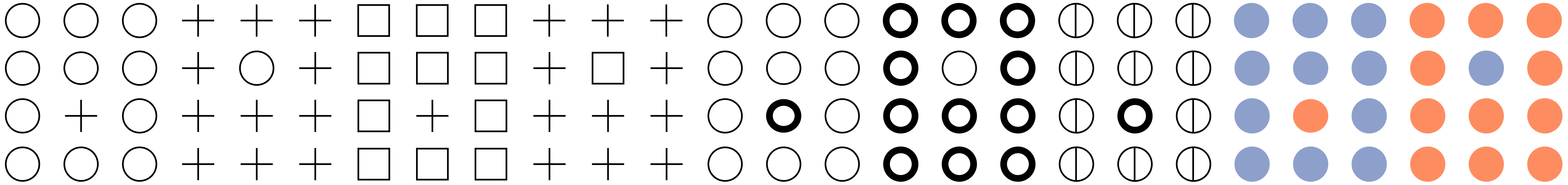
There will be cases where the number of data points and their mixing prevents you from effectively perceiving each category as a group.

In this case, we need to choose shapes that are easily distinguishable, so that local boundaries between small clusters of similar shapes can be easily seen.

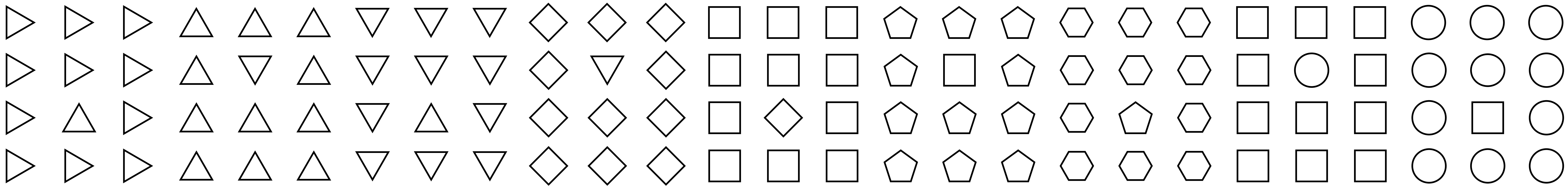
Think about the alphabet. If you encoded letters using C, D, O, and Q, because all these letter shapes are similar, it would be easy to miss a D outlier in the middle of a bunch of C’s O’s and Q’s.

Which letters would make good distinguishable symbols? For example, K, H, W and Q.

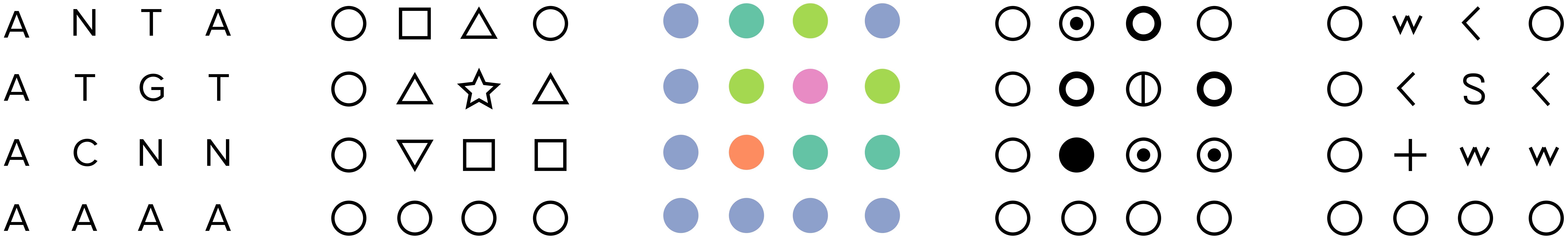
Strong visual boundaries



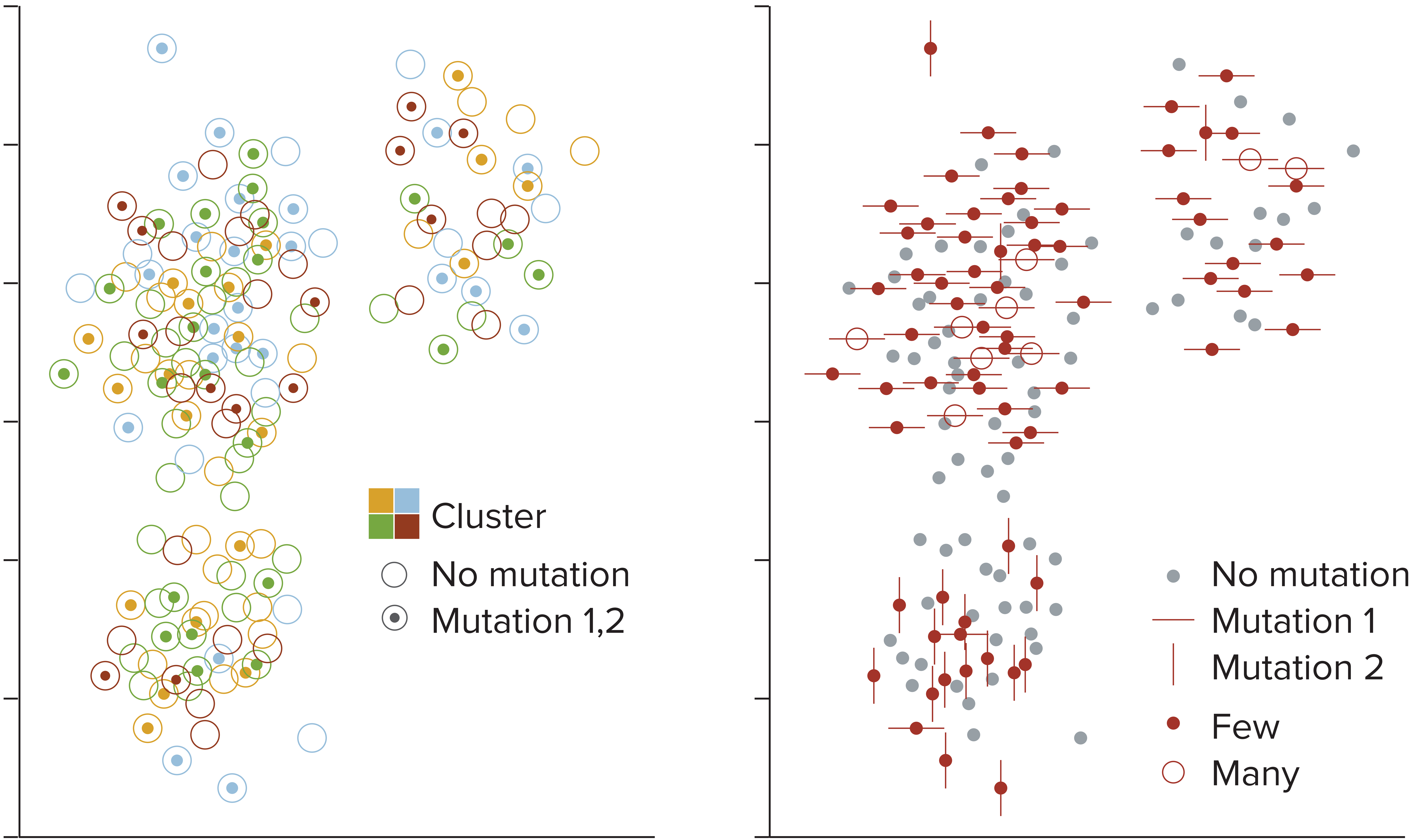
Weak visual boundaries



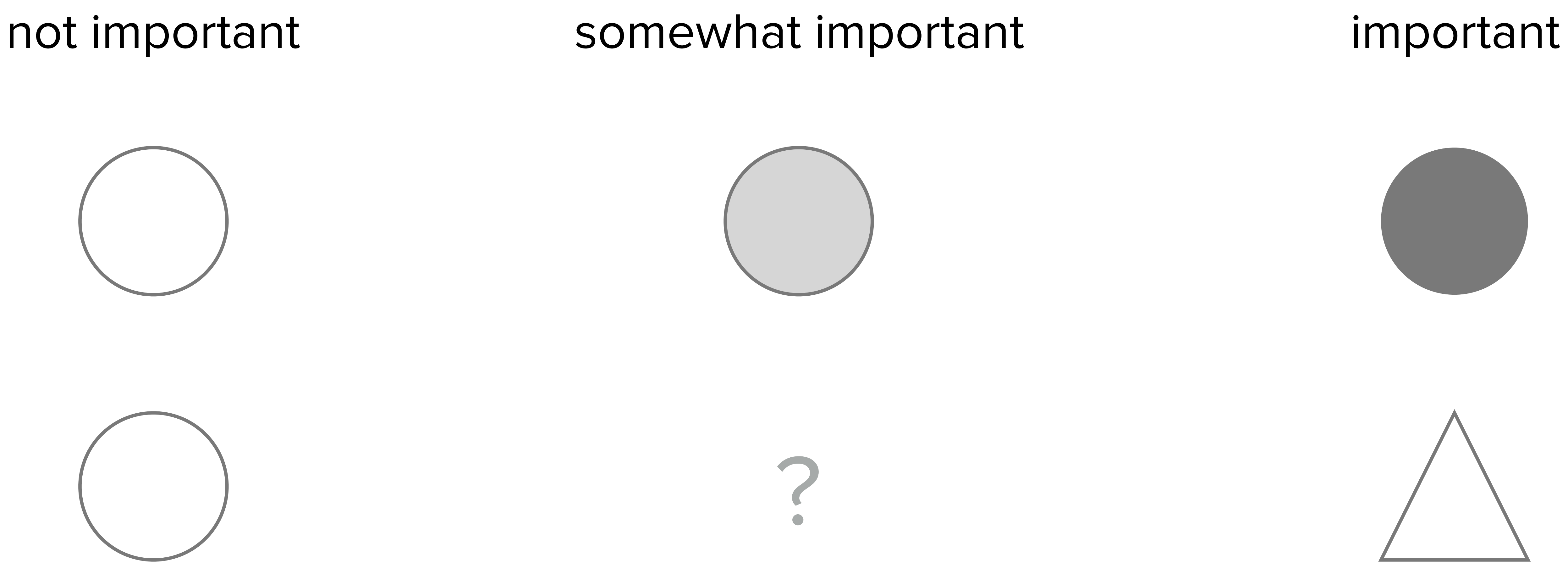
Visual boundaries of various shapes



Nat Methods (2013) 10:451.



Nat Methods (2010) 7:773.



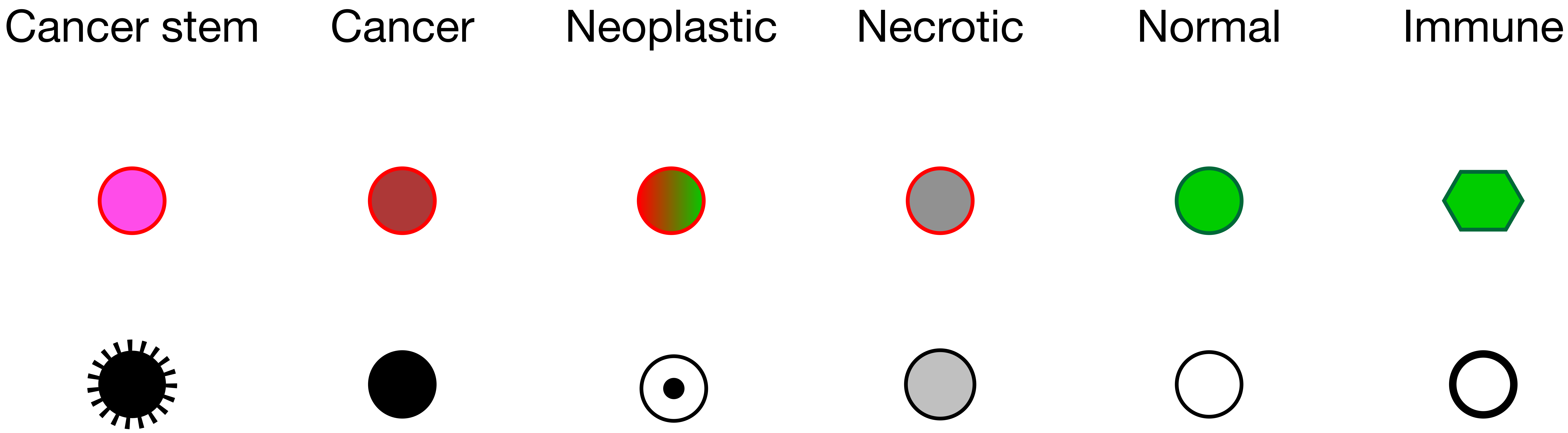
not important

somewhat important

important

?

CELL TYPES



Cancer stem

Cancer

Neoplastic

Necrotic

Normal

Immune

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

Shapes and glyphs are really important. They make up the heart of a lot of data plots.

Your default should be the circle. Be wary of trying to encode some kind of magnitude with the size of the circle—remember the point about judging areas I've made in another video.

If you need different shapes, try to map the classes as intuitively as possible onto the shapes. Adding a reagent? Use a solid shape, with hollow ones as the control. Removing something? Use a hollow shape.

If you have multiple categories, don't be afraid to experiment composing two shapes together—like a circle and a line. This isn't done often enough. If your plot has more than six or seven categories, consider presenting the data in several panels with each showing a few data categories—a technique known as small multiples.

Always attempt to map salience to relevance by using shapes with greater visual weight (fill and/or color) to distinguish and elevate important data. The use of a single color is effective at isolating a single variable. Use less prominent symbols for data that are less relevant (such as reference data included for context).

created by

Martin Krzywinski, Kim Bell-Anderson & Philip Poronnik

written and designed by

Martin Krzywinski

production

One Ski Digital Media Productions

with financial support by

University of Sydney

filmed at

University of Sydney, Australia

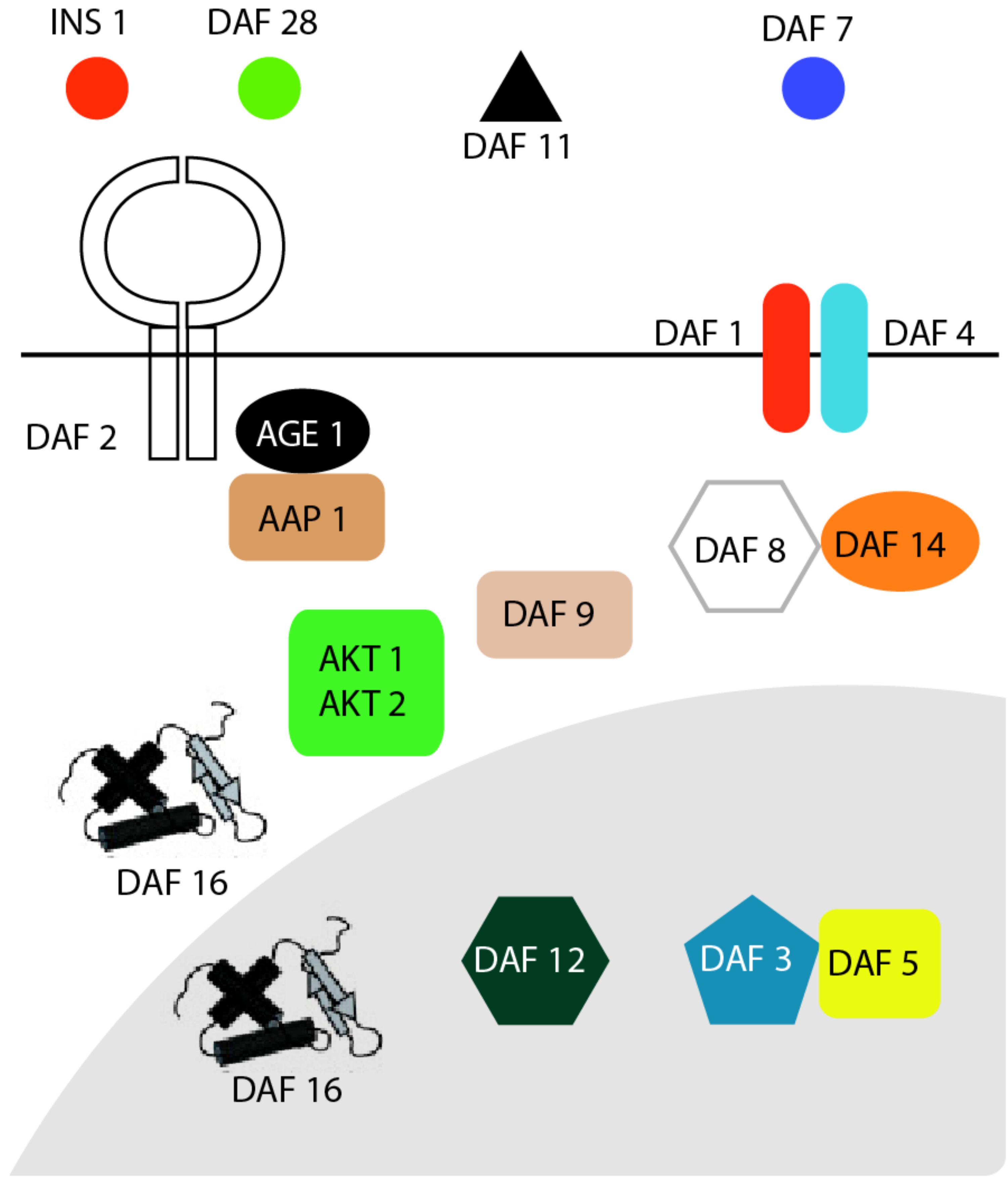
EXERCISE 1

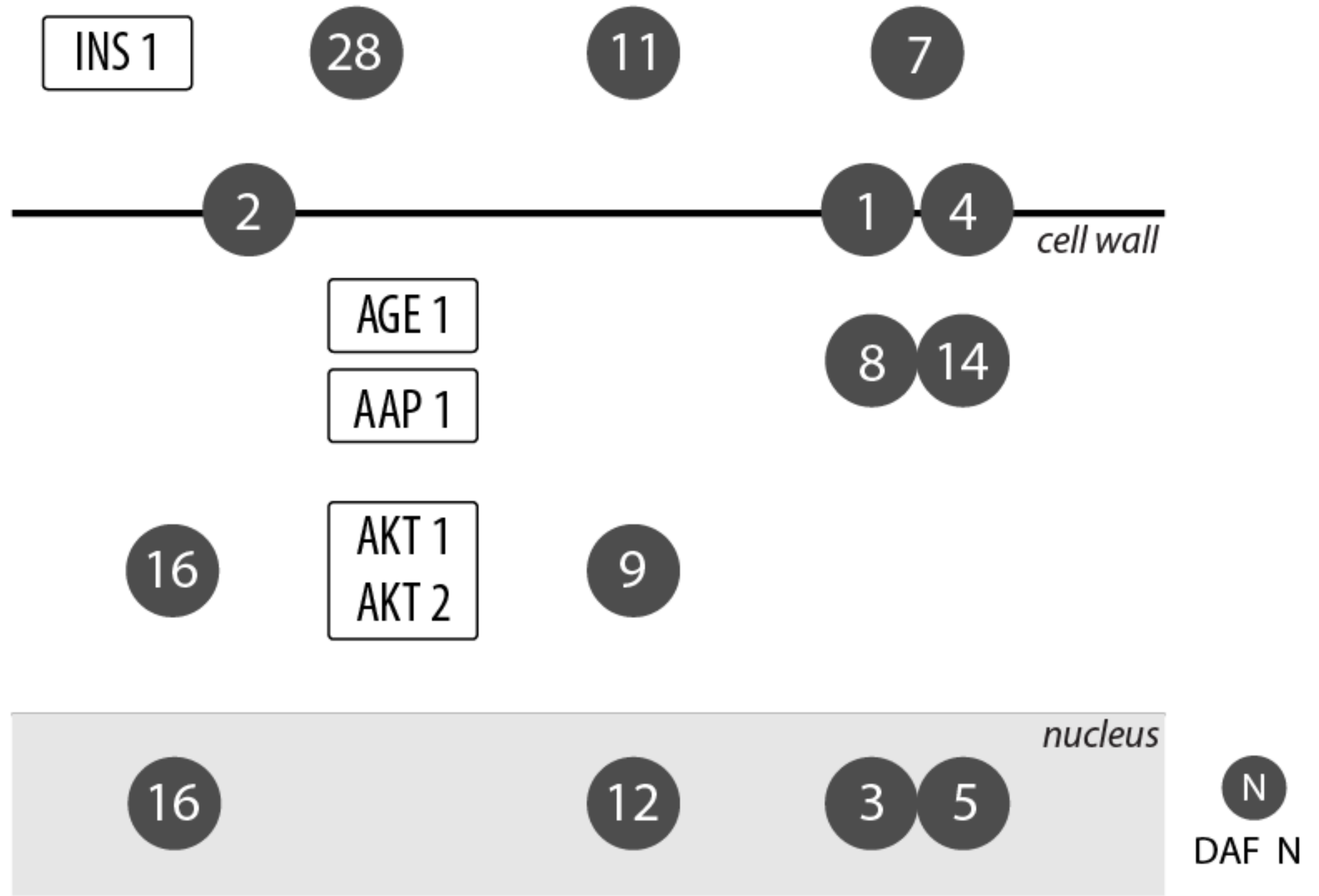
Reproduce the monkey G/H scatter plot. What happens when you move the control and inactivation data points for the same monkey closer together? At what point does grouping by spatial proximity override grouping by color similarity? Does it ever?

EXERCISE 2

Look up the paper for the figure on the top left. Try to figure out what the different shapes and color encode. What are some of the ways in which the shapes differ? Are there any shapes that suggest functional similarity? Can you find evidence of this in the paper?

Take look at my redesign at the bottom. What's your opinion about leaving the "DAF" out of the gene name and only keeping the number? Does this break naming convention? Is it easier to understand? What's the benefit and what is the cost of this design choice?





Genome Bio (2007) 8:R135.

EXERCISE 3

Suppose you are drawing mutation positions along a gene model. Assume that the line that represents the gene is the width of the screen (say 1,000 pixels) and you expect anywhere from 10 to 100–200 mutations along it.

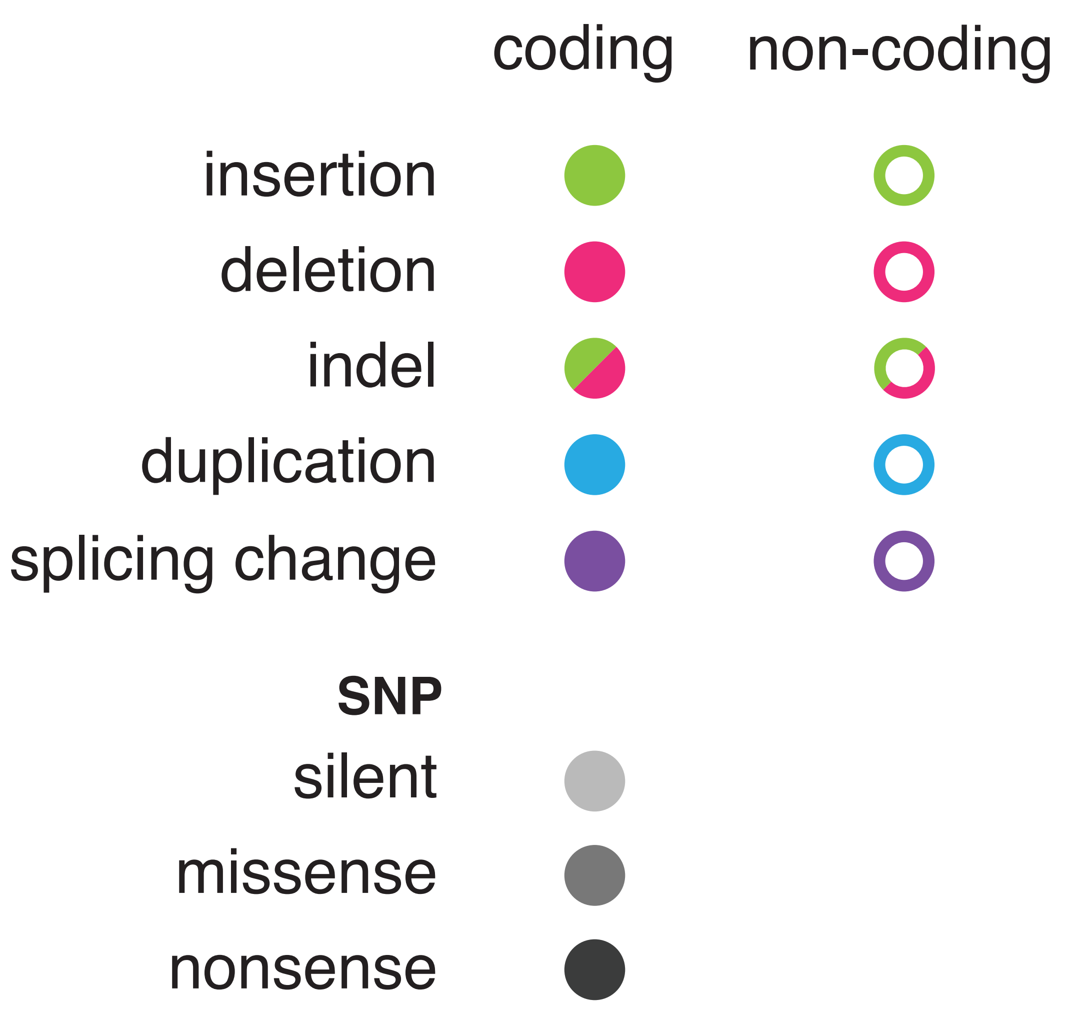
Consider the color and shape key shown here.

Redesign two versions: one color and one black and white. Remember that you can mix black/grey shapes with ones in color to easily create two groups.

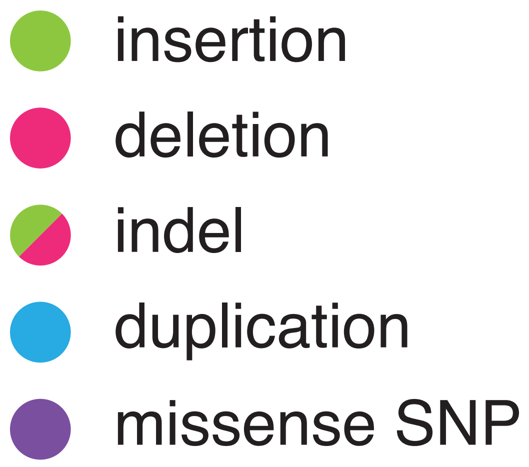
Can you present the key as a table?

🔴	duplication (noncoding)
🟢	insertion (noncoding)
🟡	indel (noncoding)
🔵	deletion (noncoding)
🟥	duplication (coding)
🟩	insertion (coding)
🟨	indel (coding)
🟦	deletion (coding)
🟠	silent SNP
🟣	missense SNP
⚫	nonsense SNP
■	splicing change

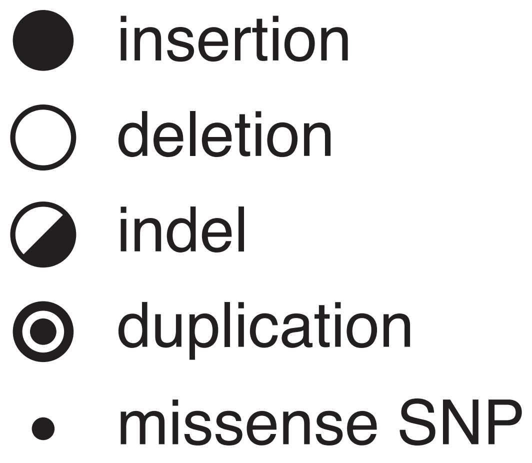
SUGGESTED SOLUTION TO EXERCISE 3



	coding	non-coding
insertion	●	○
deletion	●	○
indel	◐	○
duplication	●	○
splicing change	●	○
SNP		
silent	●	
missense	●	
nonsense	●	



●	insertion
●	deletion
◐	indel
●	duplication
●	missense SNP



●	insertion
○	deletion
◐	indel
◉	duplication
•	missense SNP

EXERCISE 4

Redesign the shape palette for SNV (single nucleotide variant), INDEL (insertion or deletion) and SNV+INDEL.

SNV	\
INDEL	X
SNV+INDEL	*