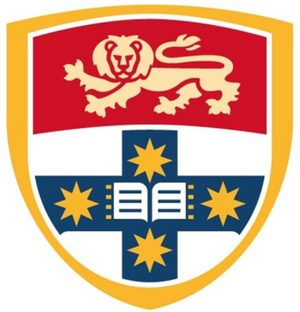
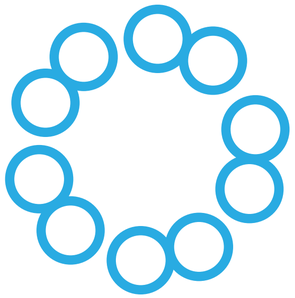


THE UNIVERSITY OF
SYDNEY



CANADA'S MICHAEL SMITH
GENOME
SCIENCES
CENTRE

ESSENTIALS OF DATA VISUALIZATION

THINKING ABOUT DRAWING DATA + COMMUNICATING SCIENCE

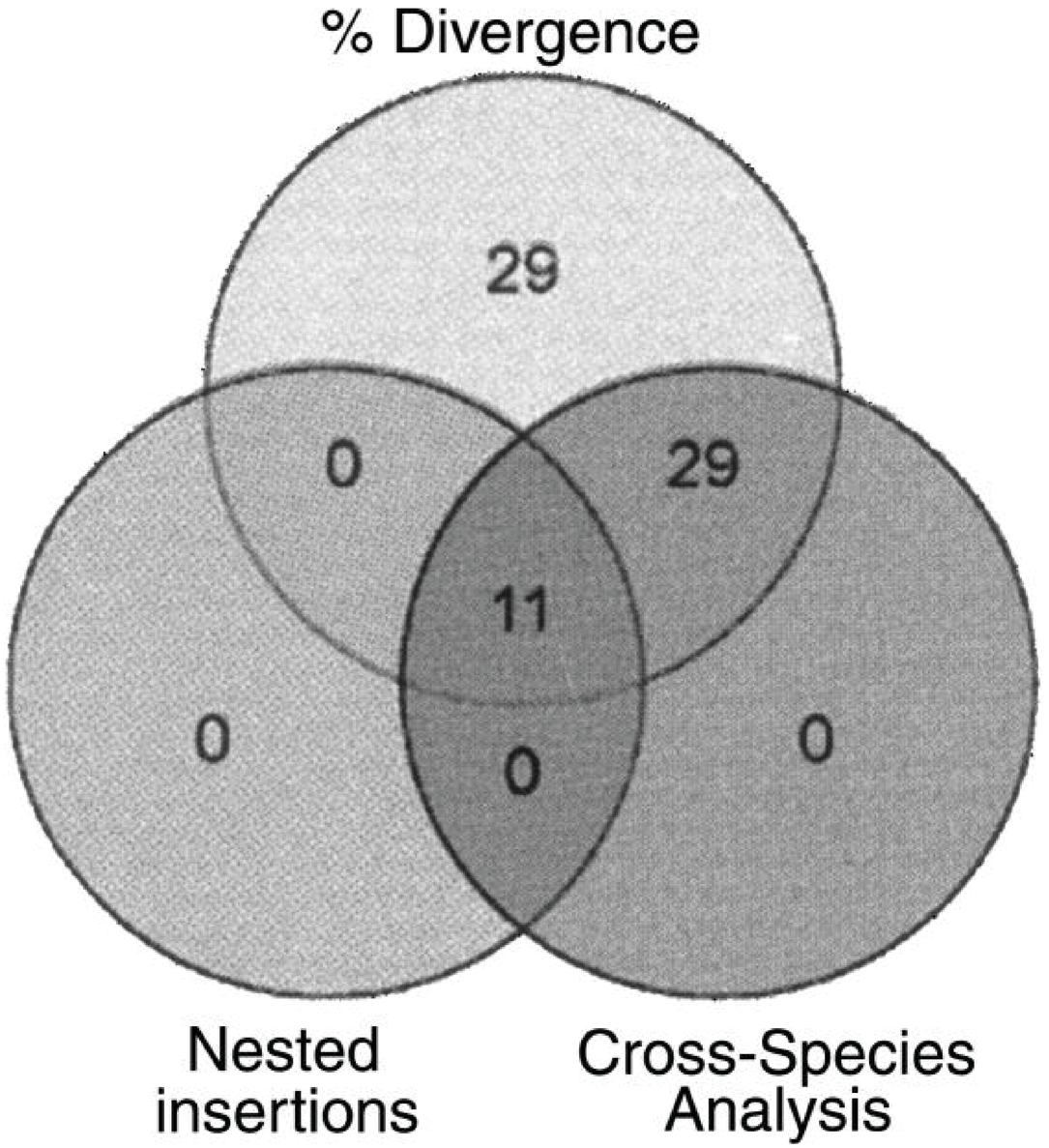
NOTHING

no data, no ink

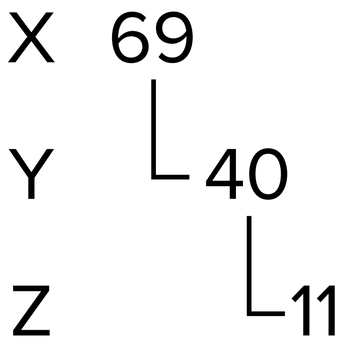
I already talked about data-to-ink-ratio. Taken to the extreme, if there is no data to show, no ink should be used.

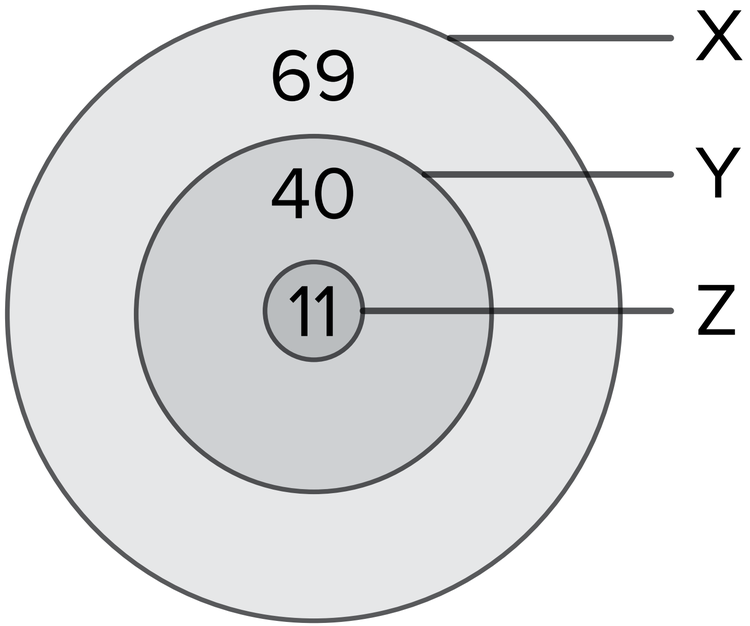
The idea of “no data to show” may correspond to a variety of scenarios. There may be sincerely no data to show—no values were collected. Or, there are no significant changes to see.

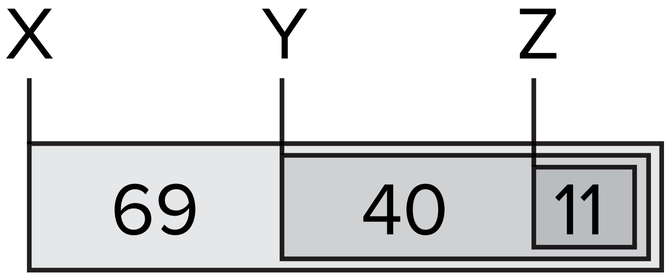
Where possible, you should use empty space to indicate lack of data or lack of change in data. You should never be distracted by something that isn’t relevant and empty space is not distracting—it really just provides contrast to adjacent elements, which presumably correspond to actual data or actionable data.



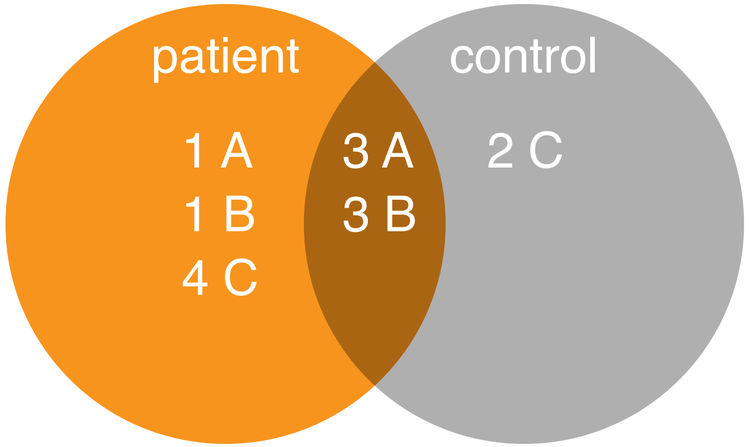
Genome Res (2007) 17:422–432.


X
69
Y
40
Z
11

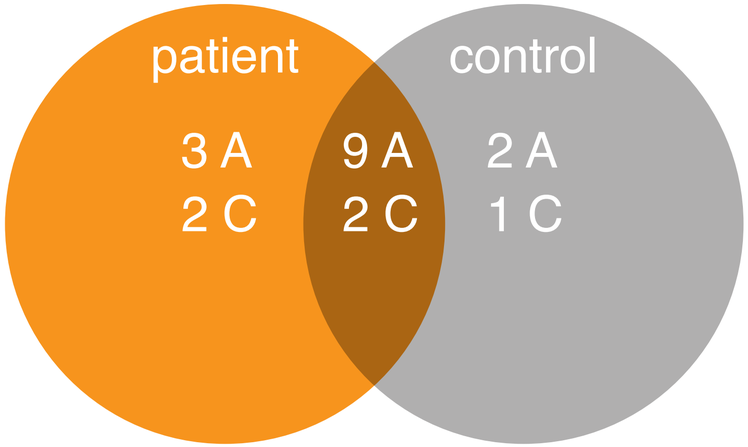

69
40
11
X
Y
Z


X
Y
Z
69
40
11

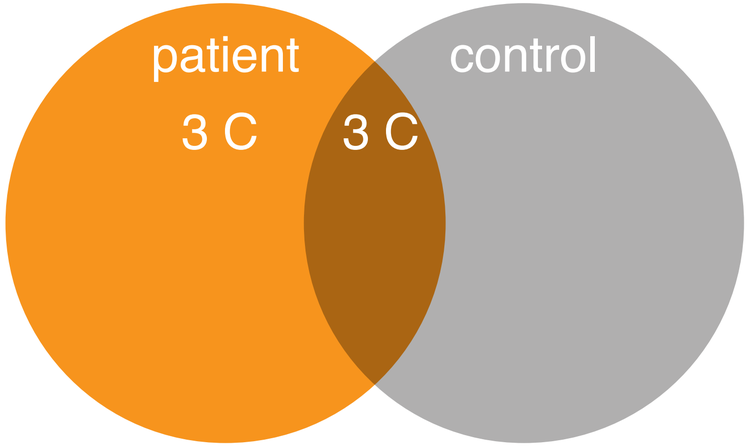
ZAP70



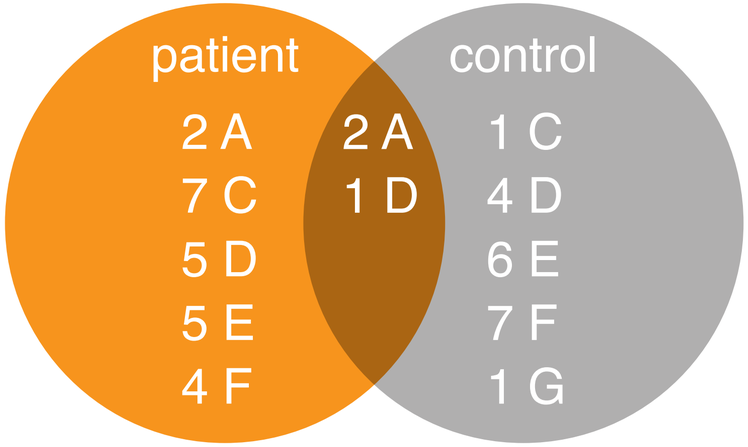
JAK2

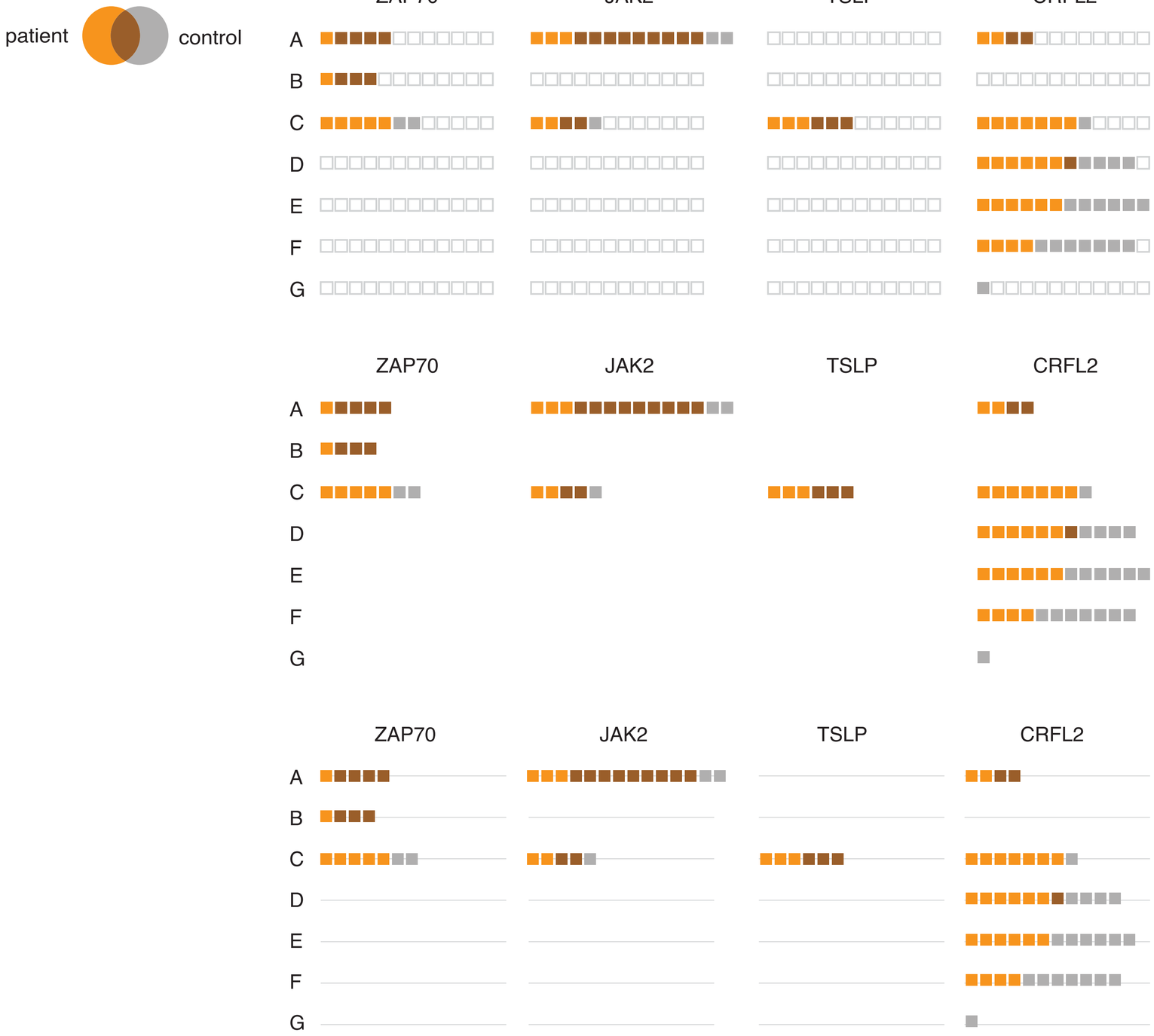


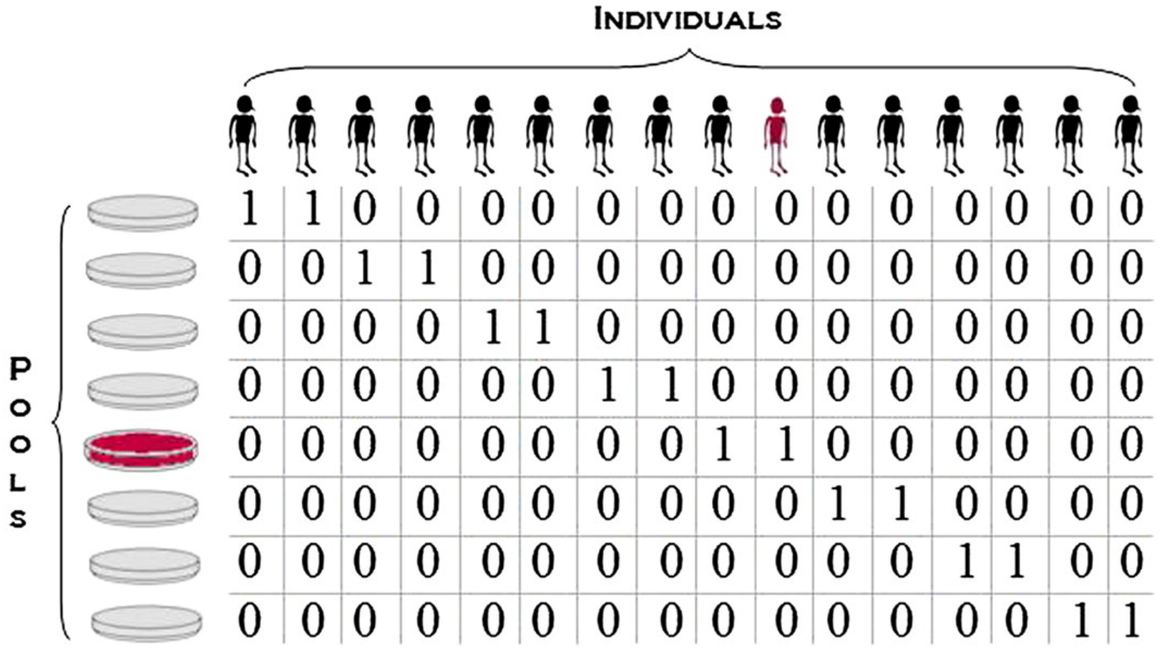
TSLP

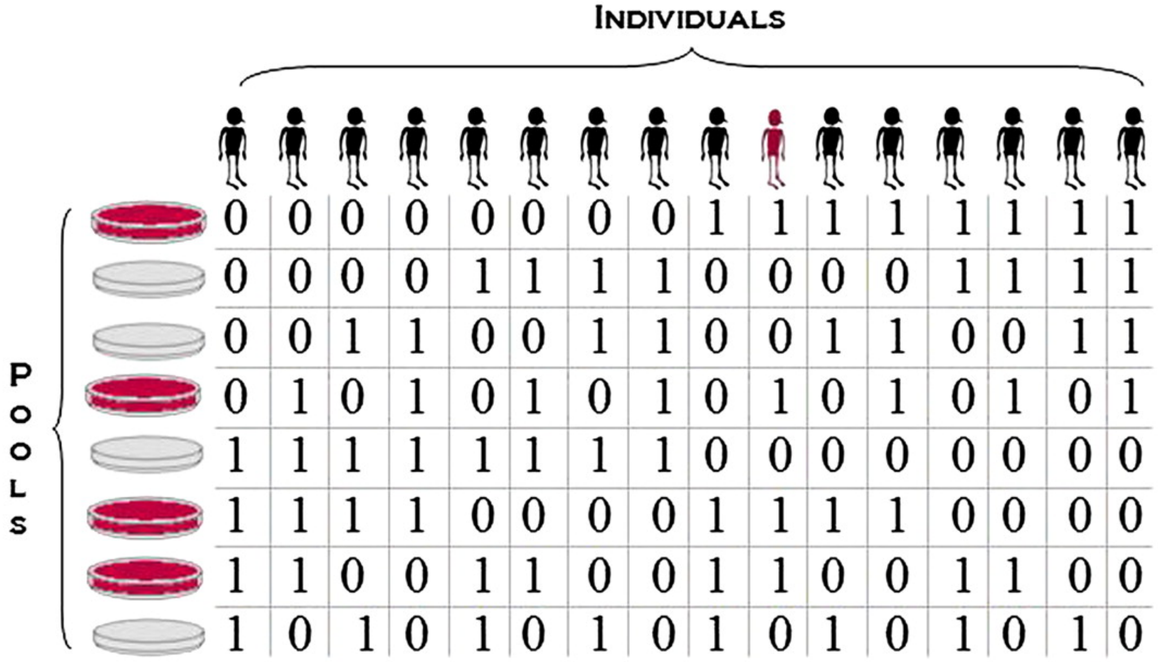


CRFL2




patient
control
ZAP70
JAK2
TSLP
CRFL2
A
B
C
D
E
F
G
ZAP70
JAK2
TSLP
CRFL2
A
B
C
D
E
F
G
ZAP70
JAK2
TSLP
CRFL2
A
B
C
D
E
F
G





Genome Res (2009) 19:1254–1261.

DISJOINT

INDIVIDUALS

POOLS

1	1	0	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	0	1	1	0	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	0	1	1	0	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	0	1	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1

OVERLAPPING

INDIVIDUALS

POOLS

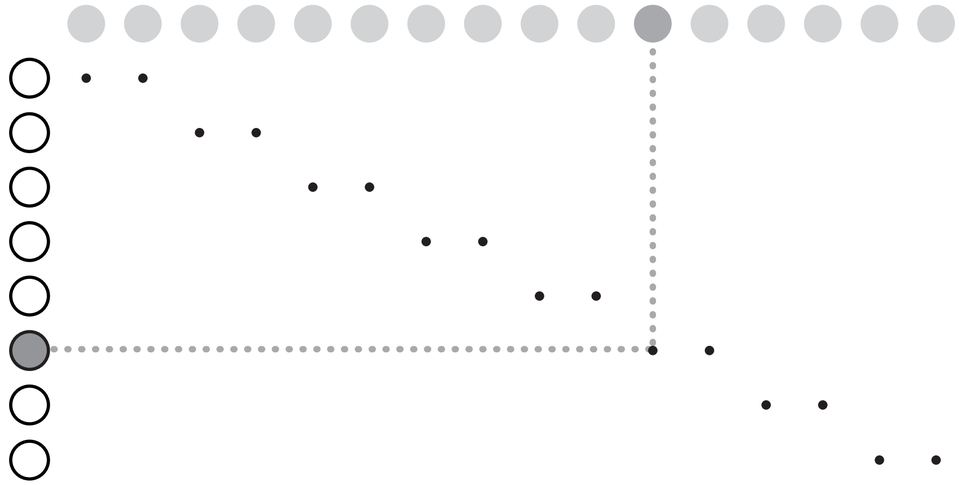
0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
0	0	1	1	0	0	1	1	0	0	1	1	0	0	0	1
0	1	0	1	0	1	0	1	0	1	0	1	0	0	1	1
1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
1	1	1	1	0	0	0	0	1	1	1	1	0	0	0	0
1	1	0	0	1	1	0	0	1	1	0	0	1	1	1	0
1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	0

DISJOINT

INDIVIDUALS

POOLS

1	1	0	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	0	1	1	0	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	0	1	1	0	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	0	1	1	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1

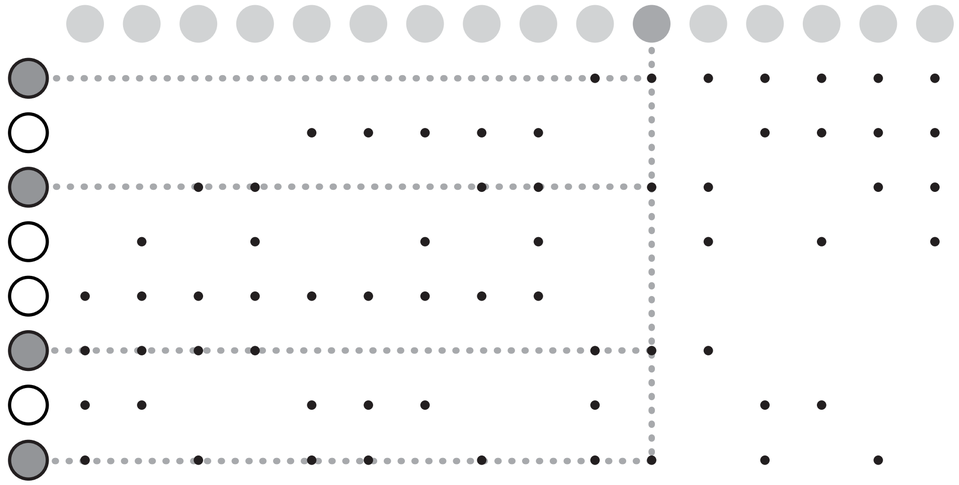


OVERLAPPING

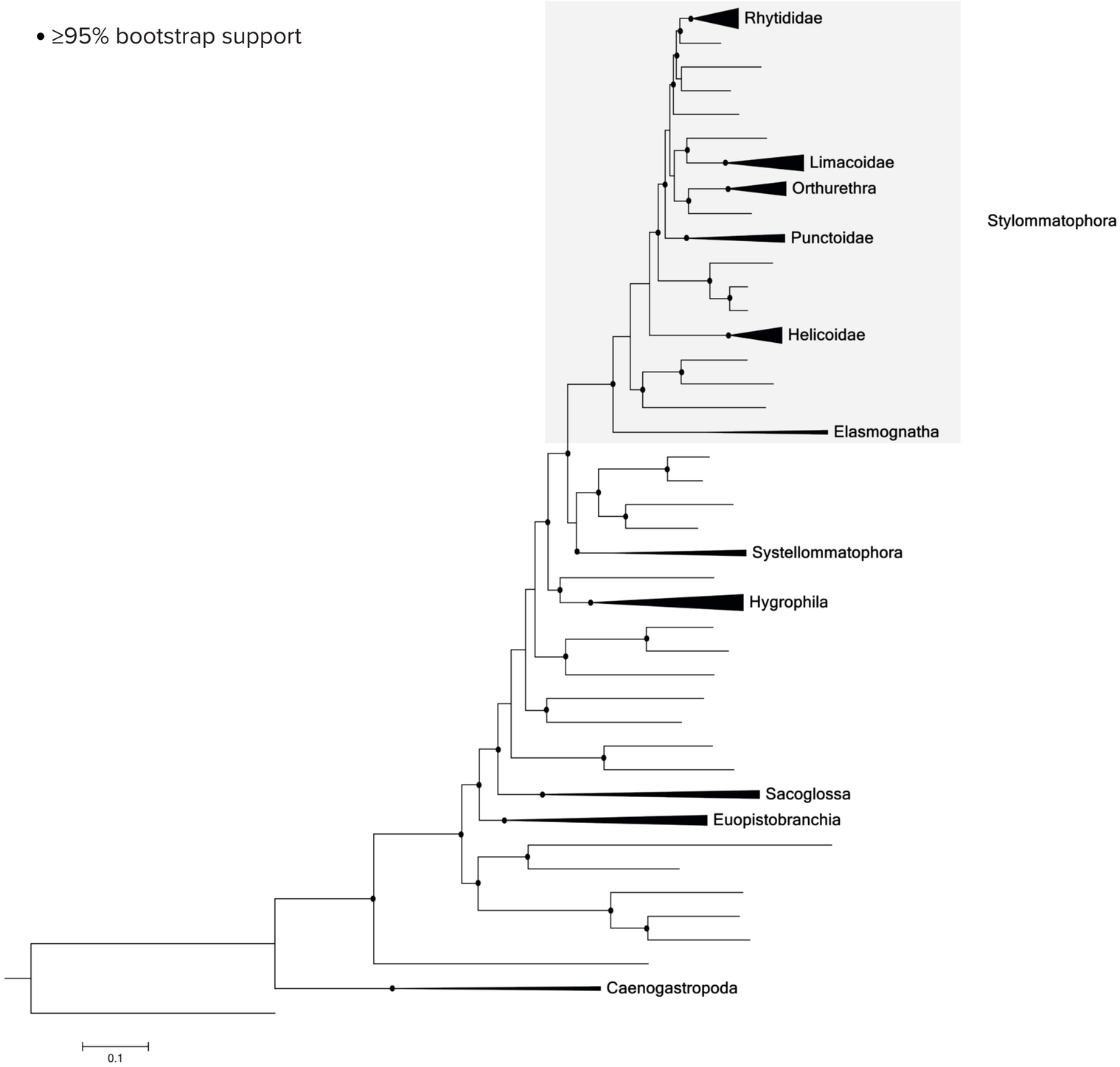
INDIVIDUALS

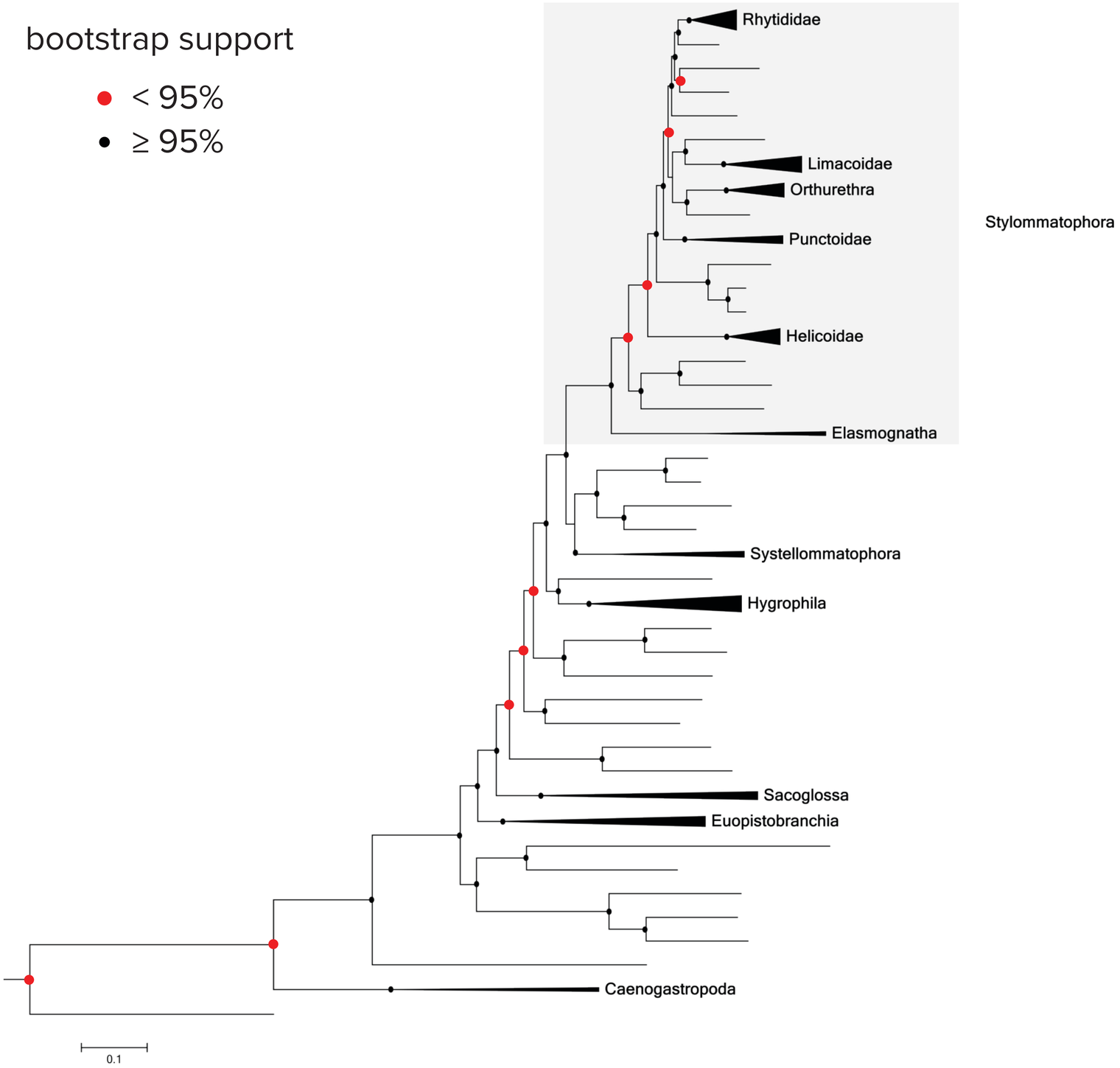
POOLS

0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
0	0	0	0	1	1	1	1	0	0	0	0	1	1	1	1
0	0	1	1	0	0	1	1	0	0	1	1	0	0	0	1
0	1	0	1	0	1	0	1	0	1	0	1	0	0	1	1
1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
1	1	1	1	0	0	0	0	1	1	1	1	0	0	0	0
1	1	0	0	1	1	0	0	1	1	0	0	1	1	1	0
1	0	1	0	1	0	1	0	1	0	1	0	1	0	0	0



- ≥95% bootstrap support




bootstrap support
< 95%
≥ 95%
Rhytididae
Limacoidae
Orthurethra
Punctoidae
Helicoidae
Elasmognatha
Stylommatophora
Systellommatophora
Hygrophila
Sacoglossa
Euopistobranchia
Caenogastropoda
0.1

The genome, say...the human genome... is an interesting thing.

Most of it is uninteresting. It's like islands in the ocean. There is a lot of ocean and we're mostly interested in the islands.

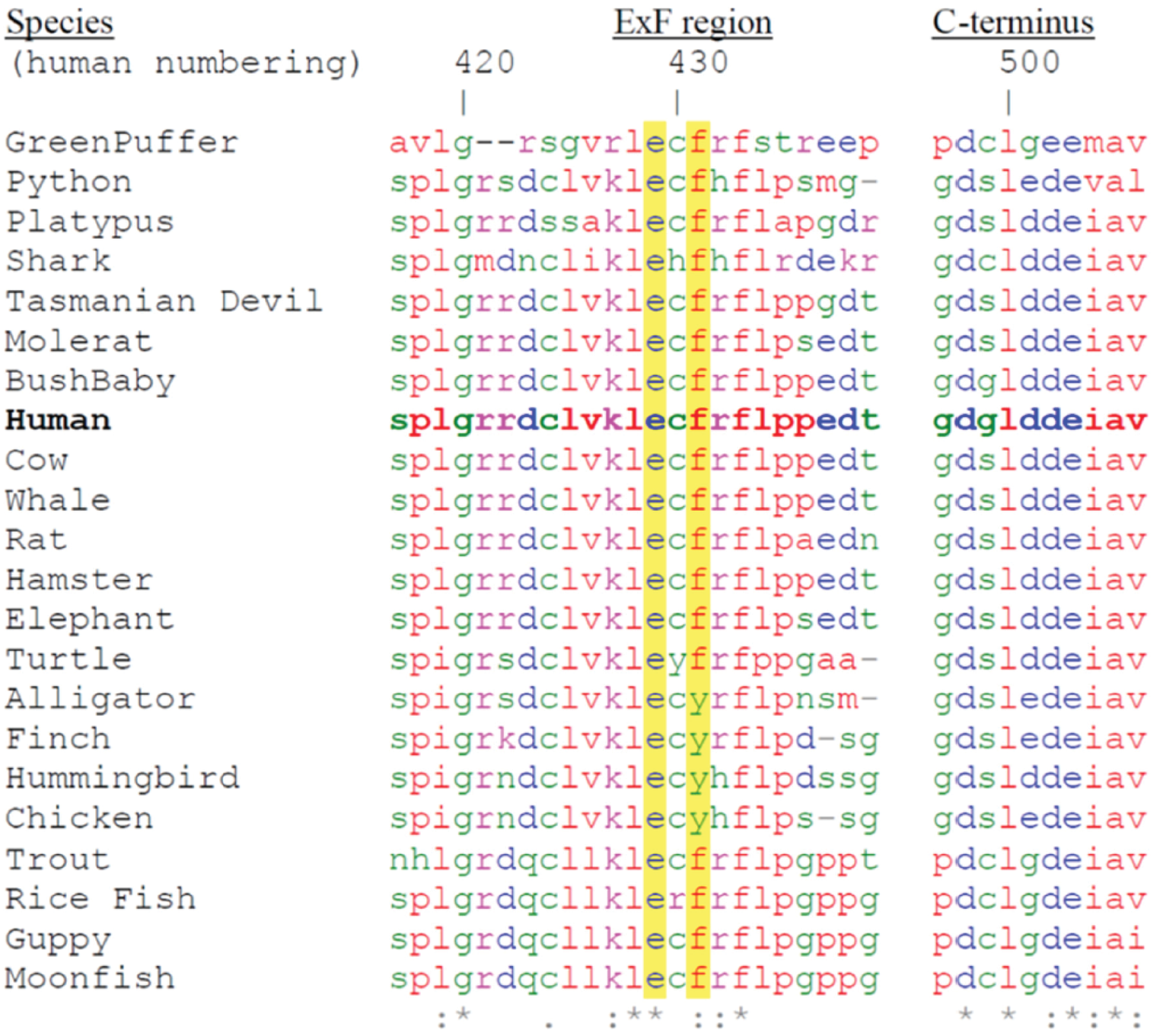
The ocean is still data, for sure, but it's not as interesting as the data from the islands. Moreover, a lot of what is interesting are the ways in which the islands are different from one another.

This means that the island data itself isn't interesting, but only how it differs from other island data.

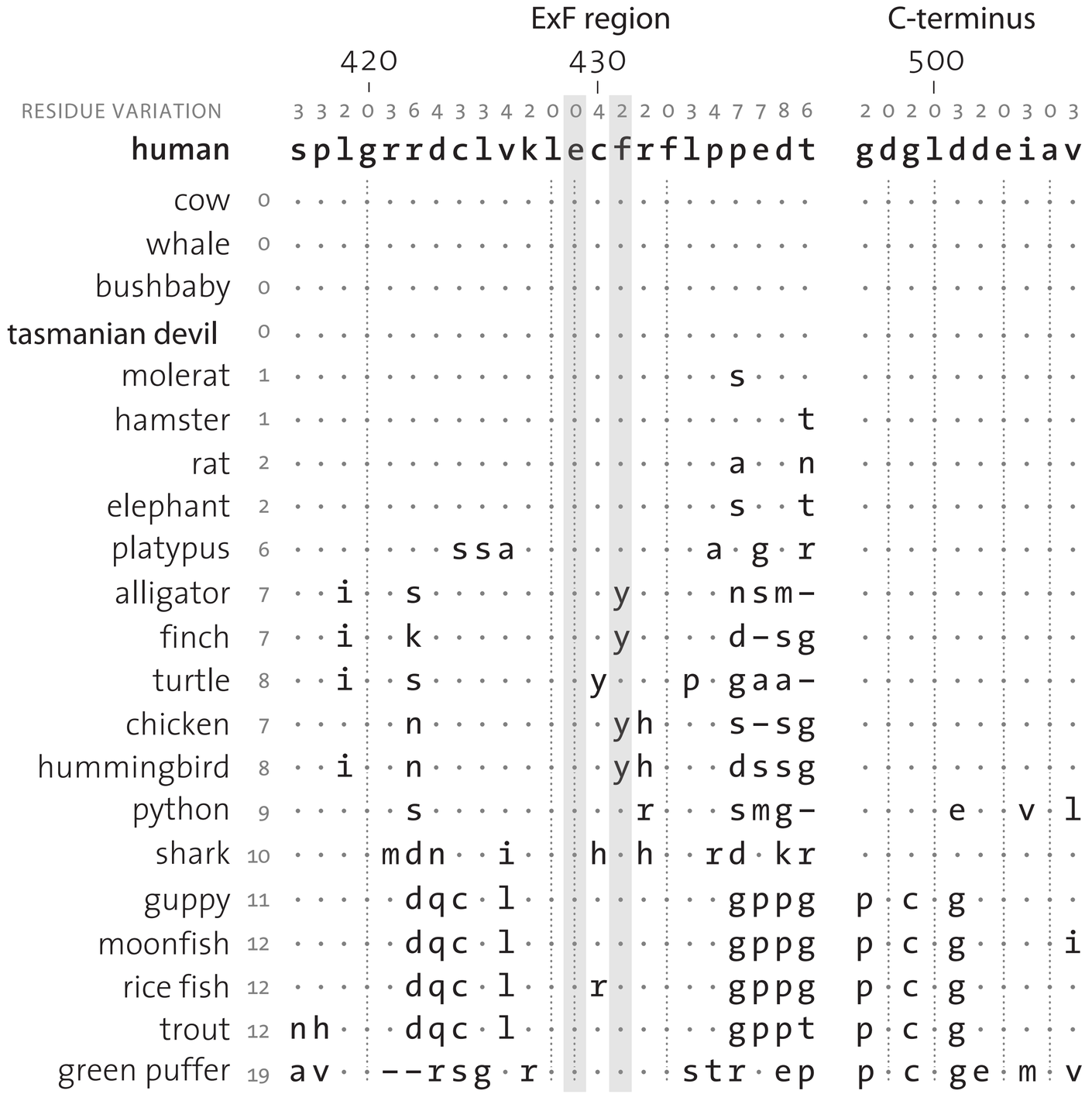
If you realize that it's the differences that are important you come to the conclusion that... most of your data should not be shown! Compute the differences and show those—everything else should disappear.

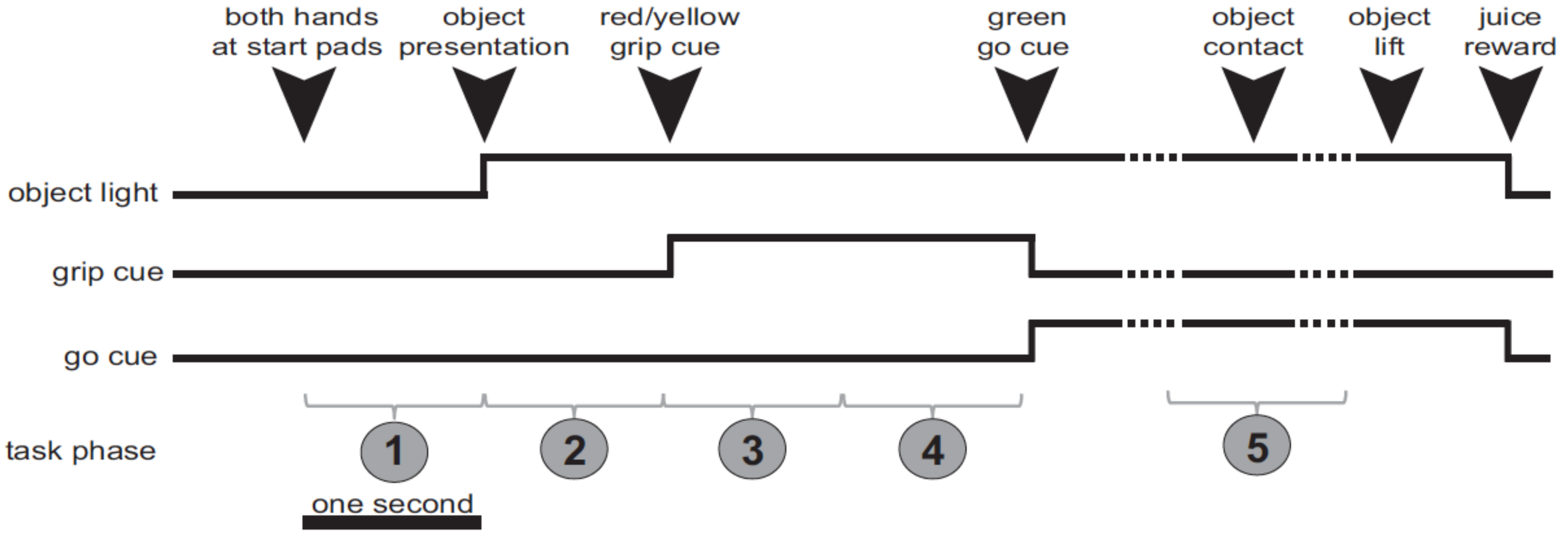
So you go from showing data... data... data... data... to nothing... nothing... oh look a difference... nothing.

Let's look at an example.

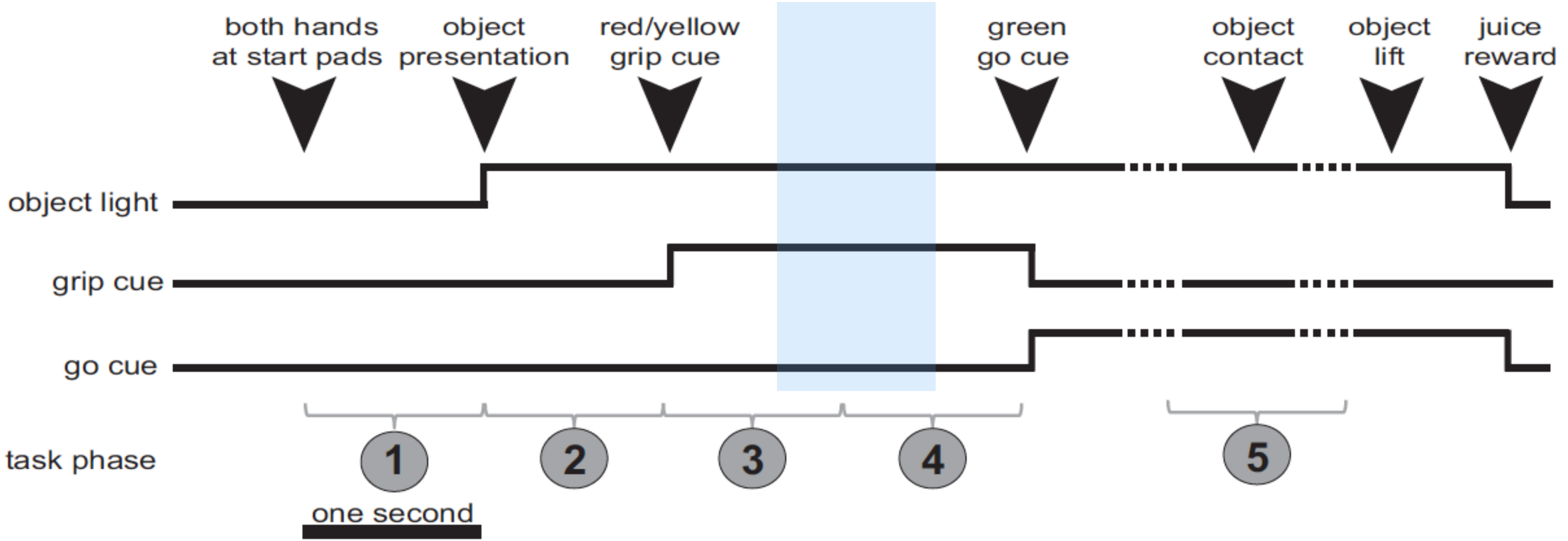


J Neurosci (2015) 35:7349–7364.

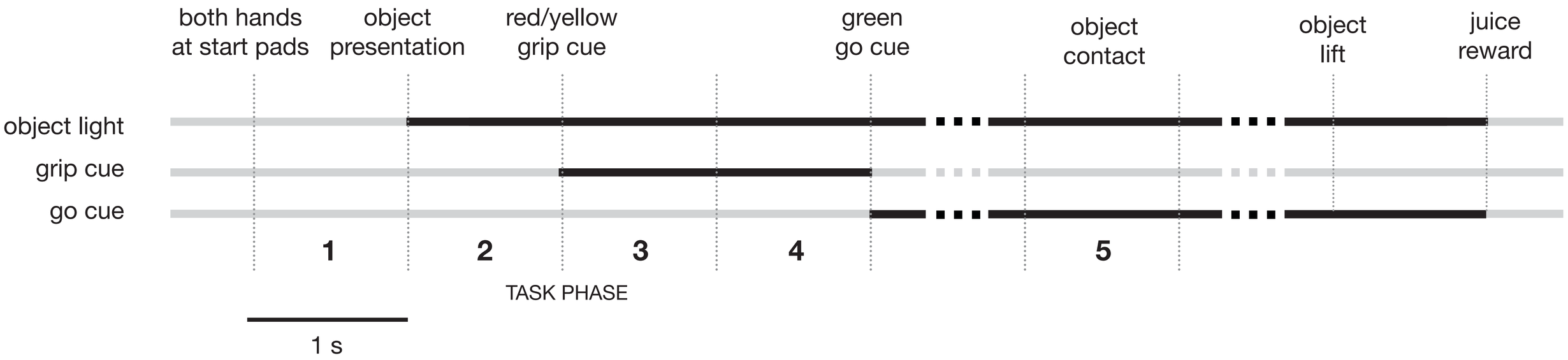

ExF region
C-terminus
420
430
500
RESIDUE VARIATION
3 3 2 0 3 6 4 3 3 4 2 0 0 4 2 2 0 3 4 7 7 8 6
2 0 2 0 3 2 0 3 0 3
human
s p l g r r d c l v k l e c f r f l p p e d t
g d g l d d e i a v
cow 0
whale 0
bushbaby 0
tasmanian devil 0
molerat 1 s
hamster 1 t
rat 2 a n
elephant 2 s t
platypus 6 s s a a g r
alligator 7 i s y n s m –
finch 7 i k y d – s g
turtle 8 i s y p g a a –
chicken 7 n y h s – s g
hummingbird 8 i n y h d s s g
python 9 s r s m g – e v l
shark 10 m d n i h h r d k r
guppy 11 d q c l g p p g p c g
moonfish 12 d q c l g p p g p c g i
rice fish 12 d q c l r g p p g p c g
trout 12 n h d q c l g p p t p c g
green puffer 19 a v – – r s g r s t r e p p c g e m v



J Neurosci (2015) 35:10888–10897.



J Neurosci (2015) 35:10888–10897.


both hands at start pads
object presentation
red/yellow grip cue
green go cue
object contact
object lift
juice reward
object light
grip cue
go cue
1
2
3
4
5
TASK PHASE
1 s

Really be sensitive to this idea of showing only differences or features that are relevant.

Look, if you have a data set and none of the observations are statistically significant, then you could argue… do you have anything to show? That's actually an interesting discussion and it comes down to what kind of conversation you're having about your data.

But if you're showing a slide for 15-30 seconds during a conference, don't bother the audience—let's assume that they're actually listening—with background noise and irrelevant outliers. Focus down on what you think means something. Showing them the things that are worth seeing, and only that.

Then, later, if they're interested, give them more.

Remember, your audience can ask for more, but it's always too late to ask for less.

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

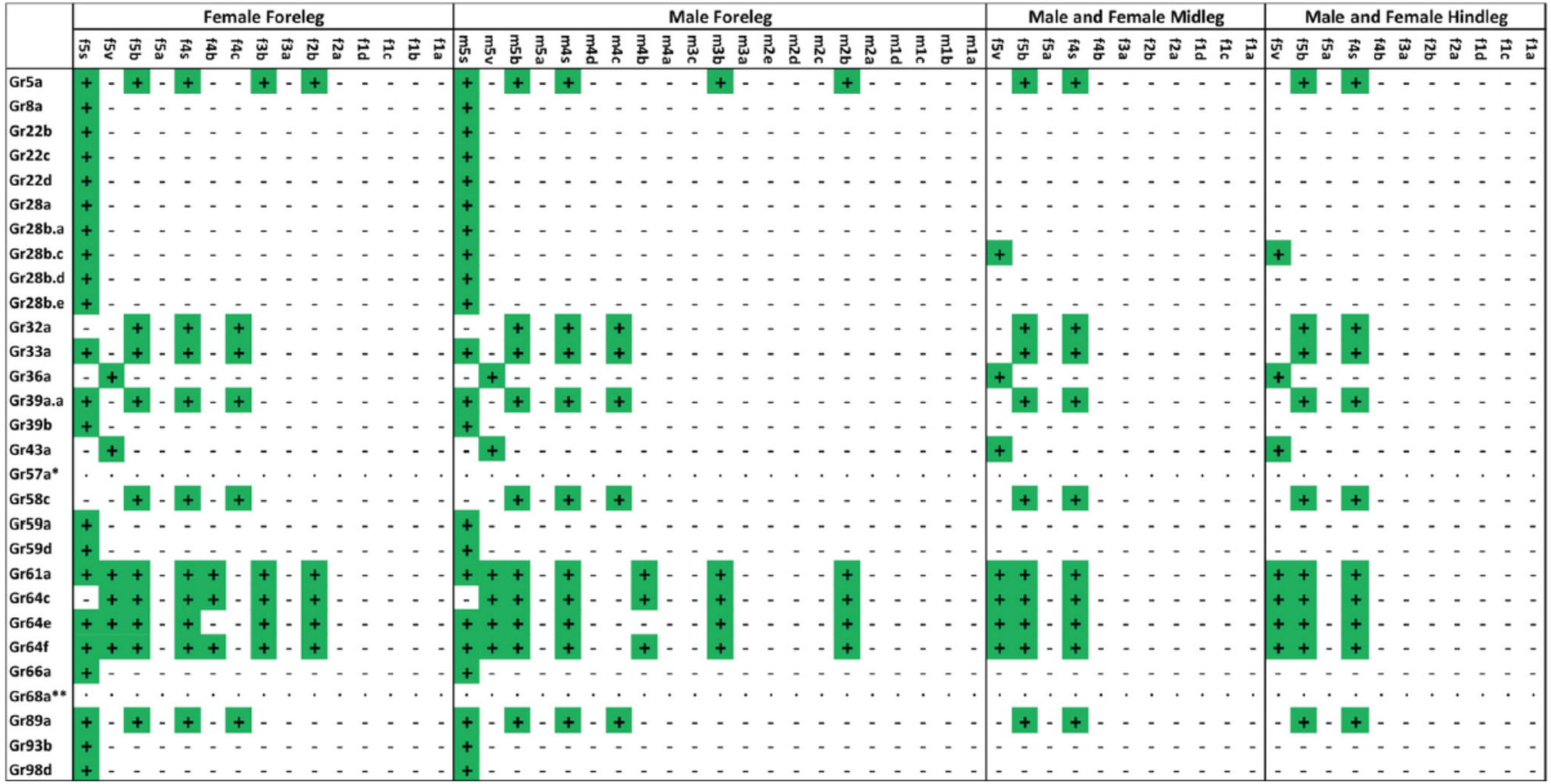
filmed at

University of Sydney, Australia

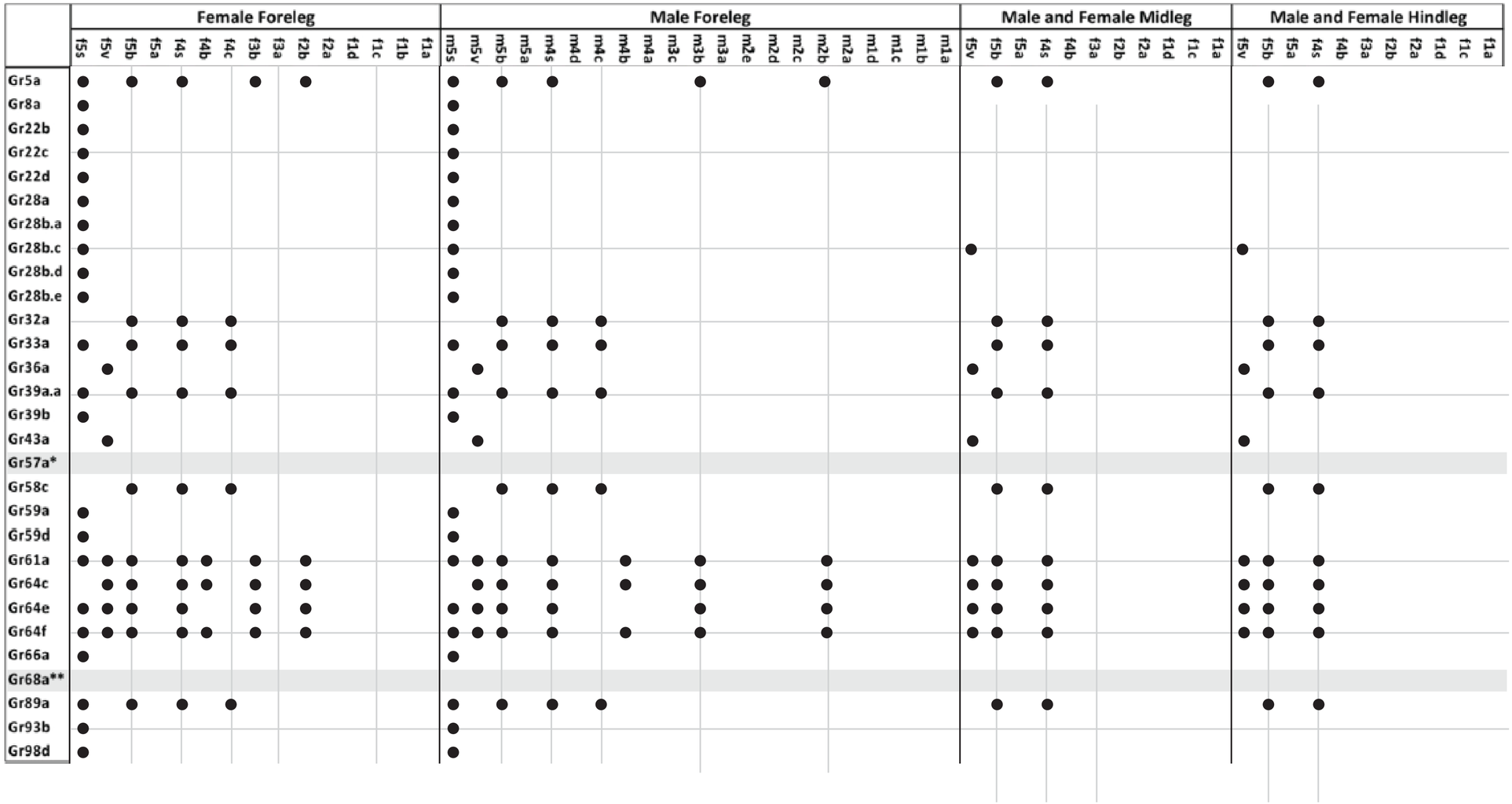
EXERCISE 1

Redesign this table.

What is the role of the green color here?



J Neurosci (2014) 34:7148–7164.


Female Foreleg
f5s f5v f5b f5a f4s f4b f4c f3b f3a f2b f2a f1d f1c f1b f1a
Male Foreleg
m5s m5v m5b m5a m4s m4d m4c m4b m4a m3c m3b m3a m2e m2d m2c m2b m2a m1d m1c m1b m1a
Male and Female Midleg
f5v f5b f5a f4s f4b f3a f2b f2a f1d f1c f1a
Male and Female Hindleg
f5v f5b f5a f4s f4b f3a f2b f2a f1d f1c f1a
Gr5a
Gr8a
Gr22b
Gr22c
Gr22d
Gr28a
Gr28b.a
Gr28b.c
Gr28b.d
Gr28b.e
Gr32a
Gr33a
Gr36a
Gr39a.a
Gr39b
Gr43a
Gr57a*
Gr58c
Gr59a
Gr59d
Gr61a
Gr64c
Gr64e
Gr64f
Gr66a
Gr68a**
Gr89a
Gr93b
Gr98d

EXERCISE 2

Redesign this table.

What is the role of the red color here?

Substrate		Samples inoculated in Tg(MoPrP^{169,170,174}) mice													
		RML control	22L control	Mock		22L			CWD			RML			Unseeded
	PMCA round			1	2	1	2	3	1	2	3	1	2	3	
Tg(MoPrP^{169,170,174})	R1	0/4	0/4	0/4	0/4	0/4	0/4	0/4	0/4	0/4	0/4	0/4	0/4	0/4	0/4
	R2	0/4	0/4	0/4	0/4	0/4	0/4	0/4	0/4	0/4	0/4	0/4	0/4	0/4	0/4
	R3	1/4	1/4	0/4	0/4	0/4	0/4	0/4	0/4	0/4	0/4	0/4	0/4	0/4	0/4
C57BL/6	R1	4/4	4/4	0/4	0/4	0/4	0/4	0/4	0/4	0/4	0/4	0/4	0/4	0/4	0/4
	R2	4/4	4/4	0/4	0/4	0/4	0/4	0/4	0/4	0/4	0/4	0/4	0/4	0/4	0/4
	R3	4/4	4/4	0/4	0/4	0/4	0/4	0/4	0/4	0/4	0/4	0/4	0/4	0/4	0/4

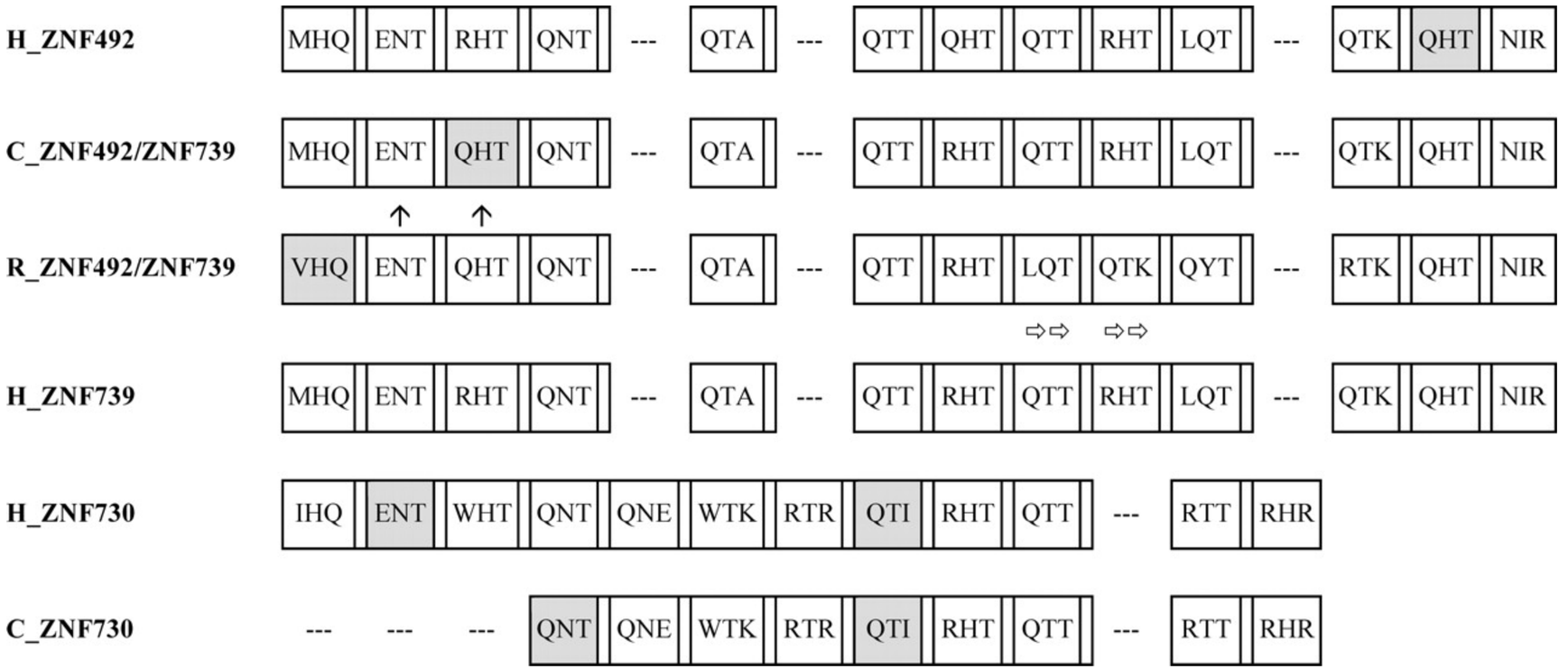
J Neurosci (2014) 34:1022–1027.

Substrate		PMCA round	RML control	22L control	Samples inoculated in Tg(MoPrP^{169,170,174}) mice: Mock 1	Mock 2	22L 1	22L 2	22L 3	CWD 1	CWD 2	CWD 3	RML 1	RML 2	RML 3	Unseeded
Tg(MoPrP^{169,170,174})		R1	.	.	.	.	.	.	.	.	.	.	.	.	.	
		R2	.	.	.	.	.	.	.	.	.	.	.	.	.	
		R3	1	4	.	.	.	.	.	.	.	.	.	.	.	
C57BL/6		R1	4	4	.	.	.	.	.	.	.	.	.	.	.	
		R2	4	4	.	.	.	.	.	.	.	.	.	.	.	
		R3	4	4	.	.	.	.	.	.	.	.	.	.	.	

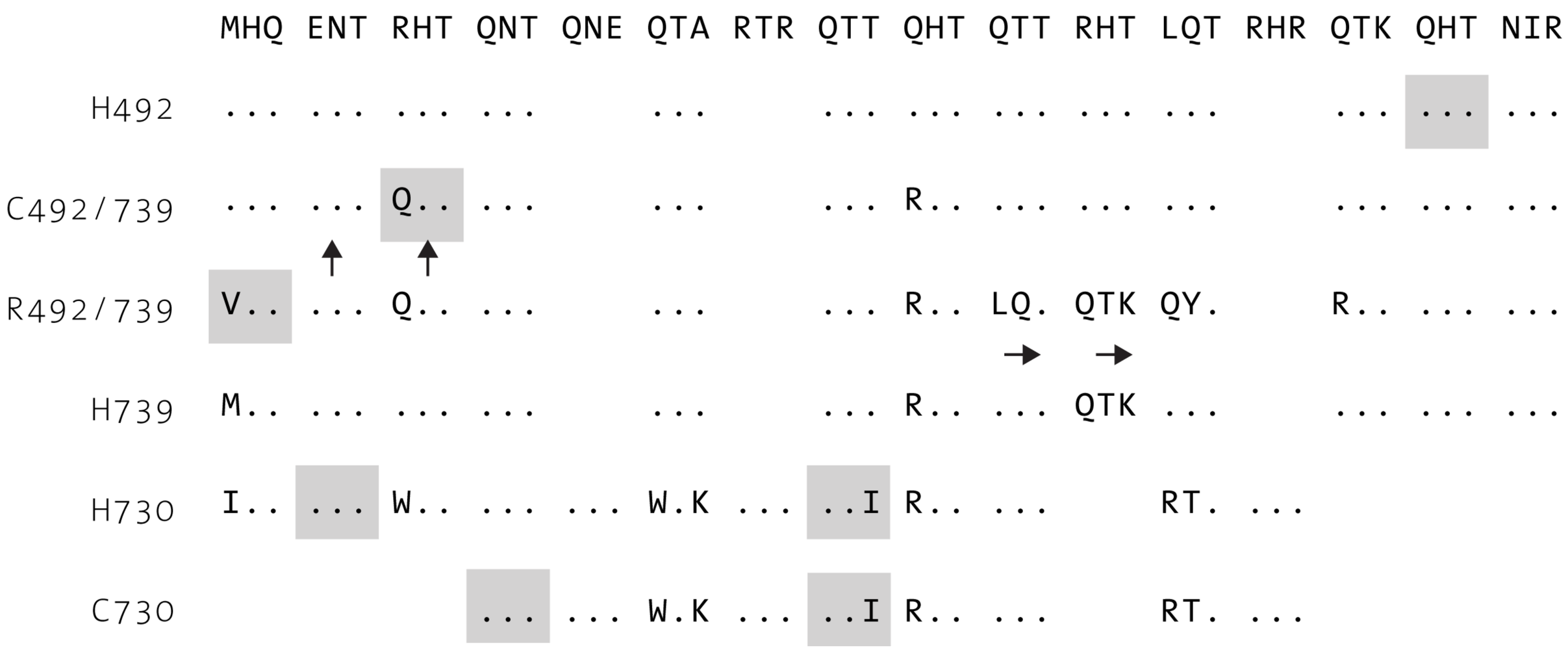
Substrate		PMCA round	control		Samples inoculated in TG mice											
			RML	22L	Mock		22L			CWD			RML			
					1	2	1	2	3	1	2	3	1	2	3	
Tg		R1	•	•	•	•	•	•	•	•	•	•	•	•	•	
		R2	•	•	•	•	•	•	•	•	•	•	•	•	•	
		R3	**1**	**1**	•	•	•	•	•	•	•	•	•	•	•	
C57BL/6		R1	**4**	**4**	•	•	•	•	•	•	•	•	•	•	•	
		R2	**4**	**4**	•	•	•	•	•	•	•	•	•	•	•	
		R3	**4**	**4**	•	•	•	•	•	•	•	•	•	•	•	

EXERCISE 3

Redesign this figure.



Genome Res (2006) 16:584–594.


MHQ ENT RHT QNT QNE QTA RTR QTT QHT QTT RHT LQT RHR QTK QHT NIR
H492 ... ...
C492/739 ... ... Q.. ... ... R.. ... ...
R492/739 V.. ... Q.. ... ... R.. LQ. QTK QY. R.. ... ...
H739 M.. ... ... R.. ... QTK ... ...
H730 I.. ... W.. ... ... W.K ... ..I R.. ... RT. ...
C730 ... ... W.K ... ..I R.. ... RT. ...