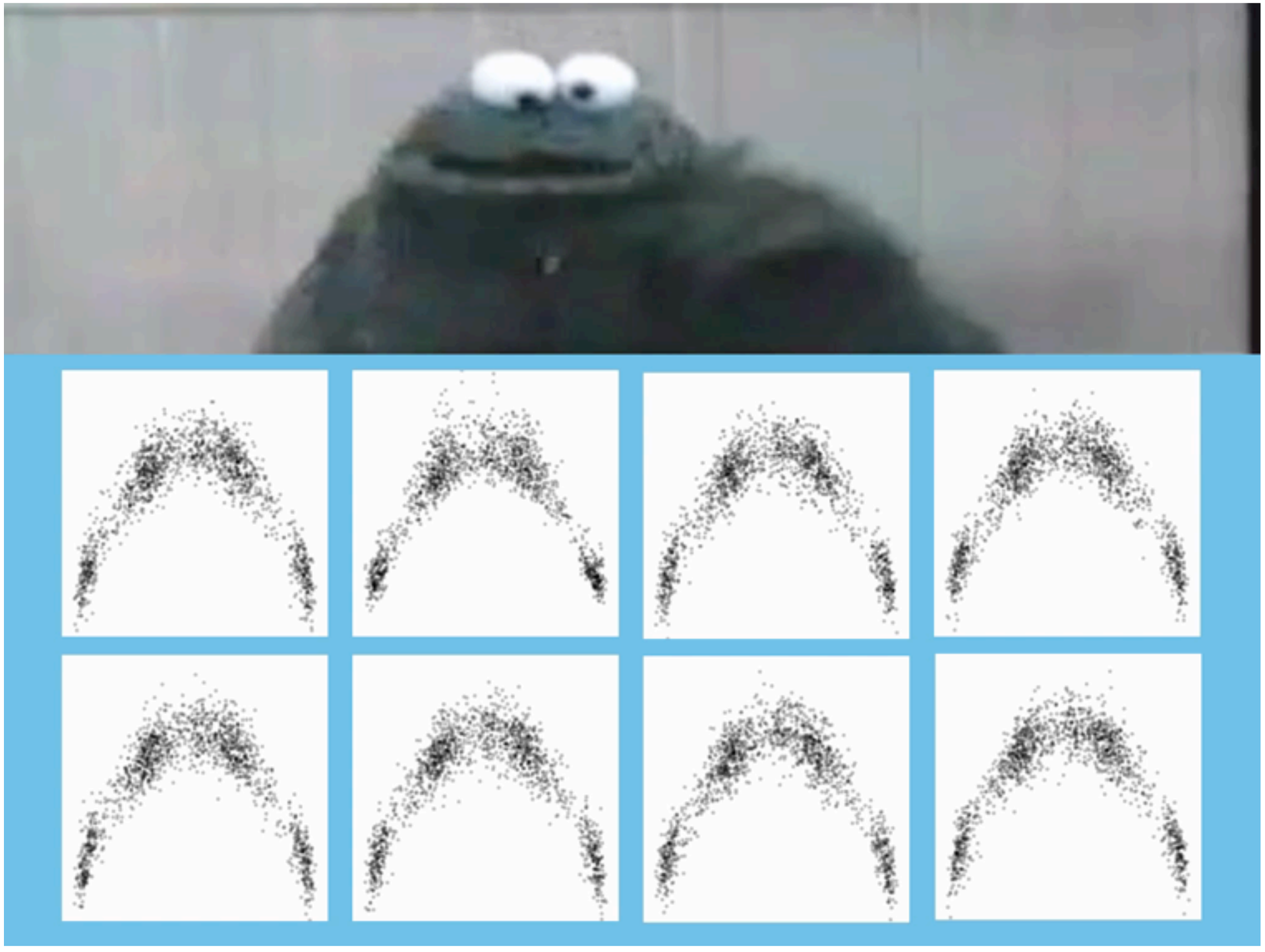
Data Visualization
Discover, Explore and be Skeptical

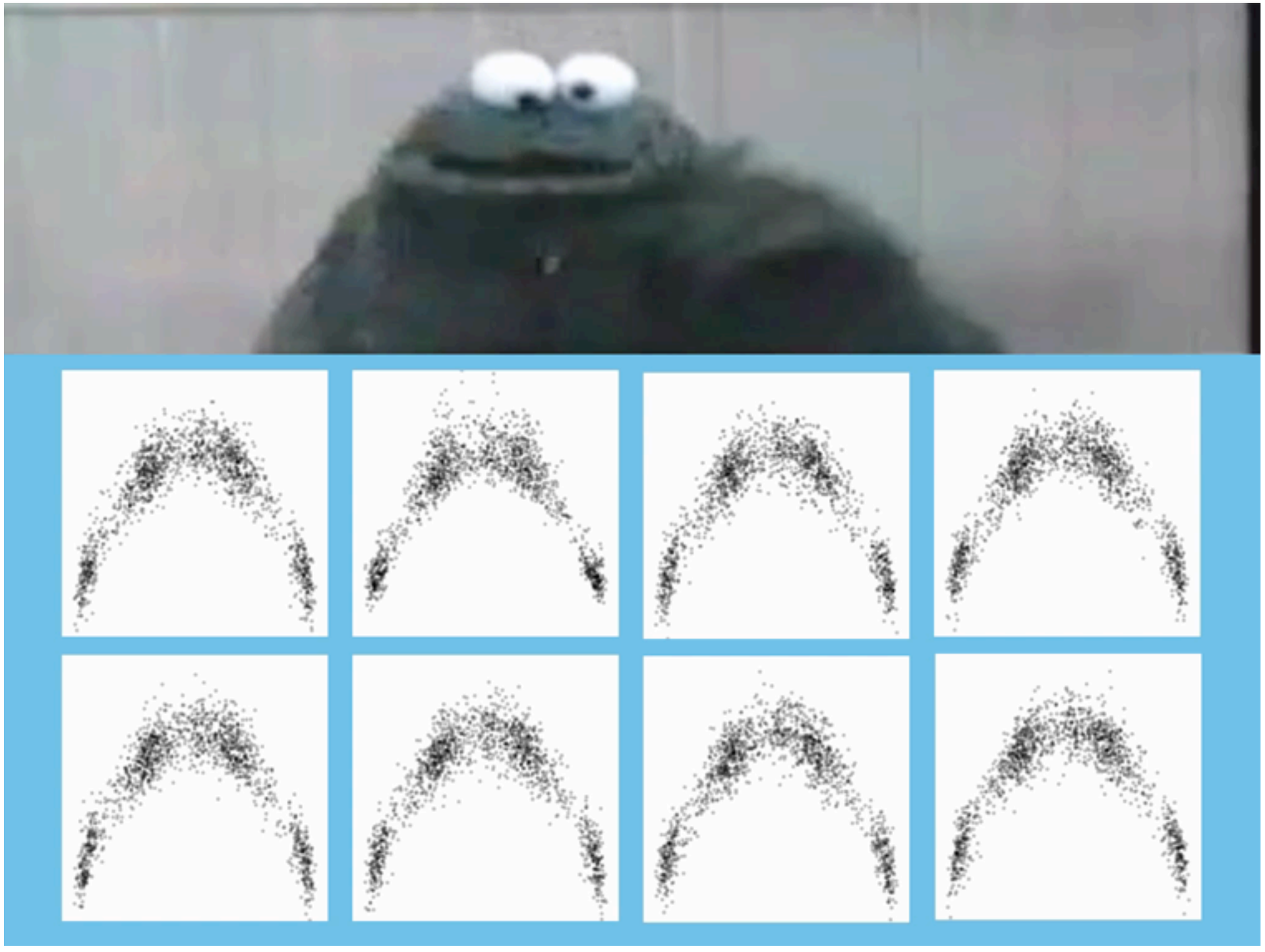
Di Cook
Statistics, Iowa State University
soon to be Business Analytics, Monash University

Seminar 3
Inference and Exploration

- Discoveries need to be calibrated by what might have been possible. Maintain a healthy skepticism.
- Underlying plots of data, are assumptions that implicitly specifying null hypotheses: what would you see if there really was nothing happening.
- Exploratory and inferential ARE NOT mutually exclusive.

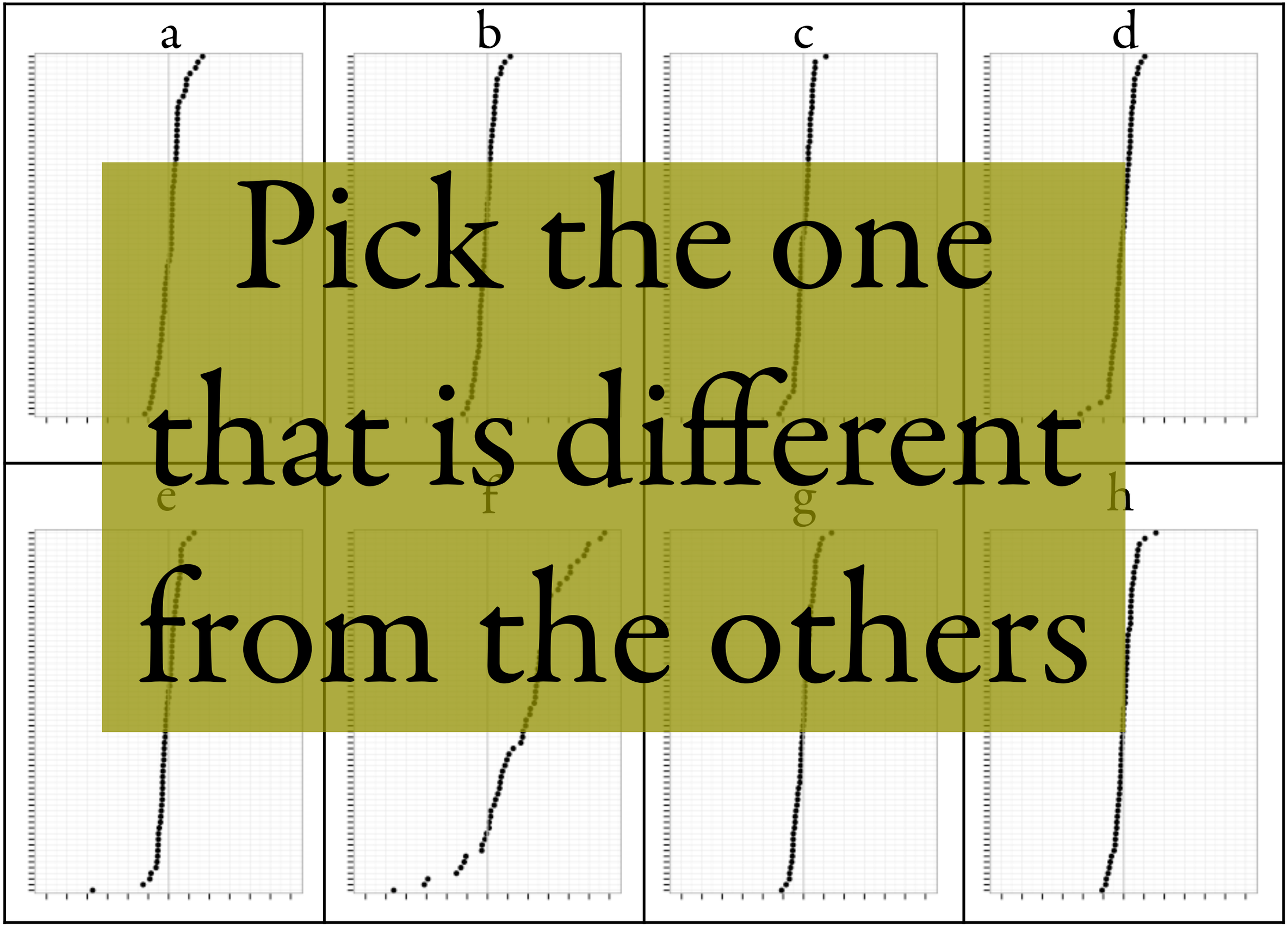


Video made by Hadley Wickham



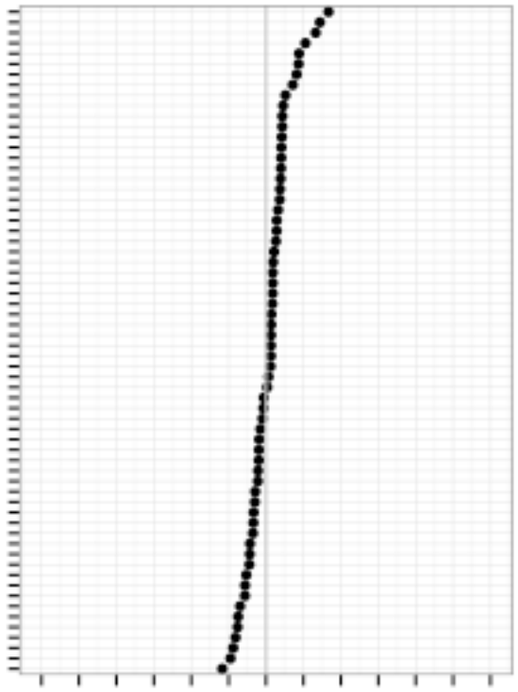
Video made by Hadley Wickham

Here is the math gap exploration placed in the CONTEXT of there being NO MATH GAP ...

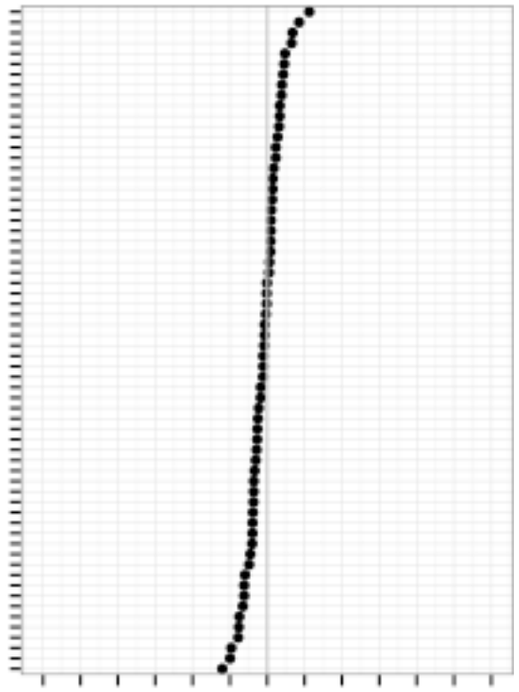


Pick the one that is different from the others

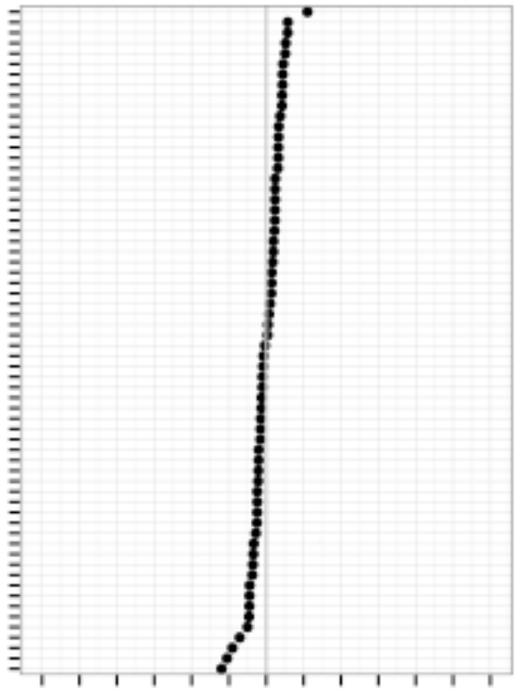
a



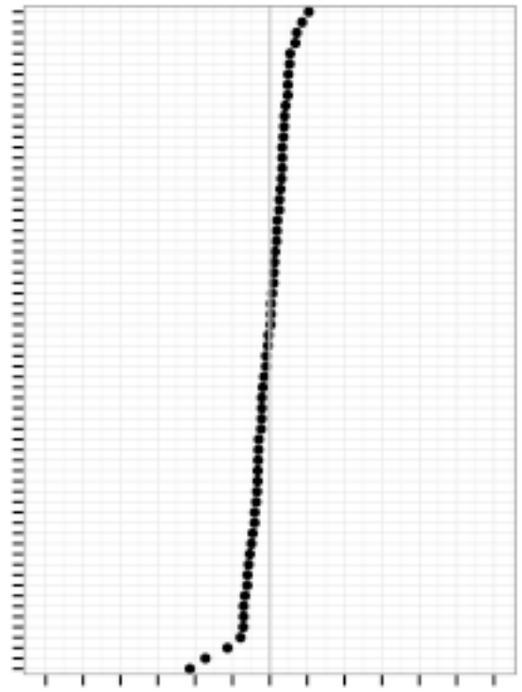
b



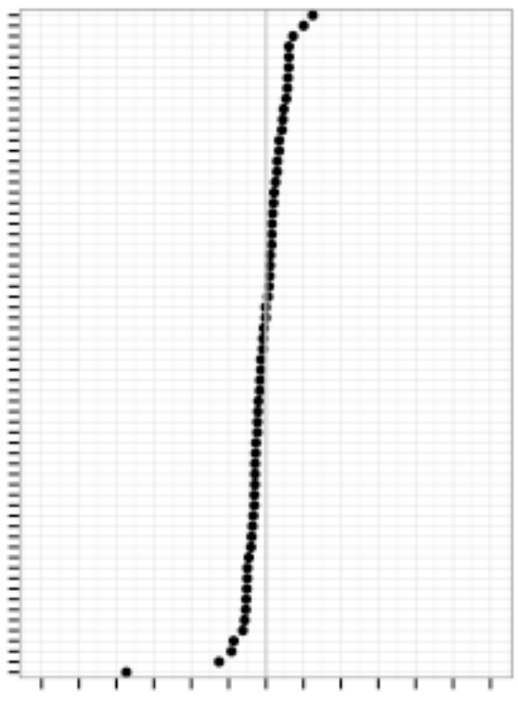
c



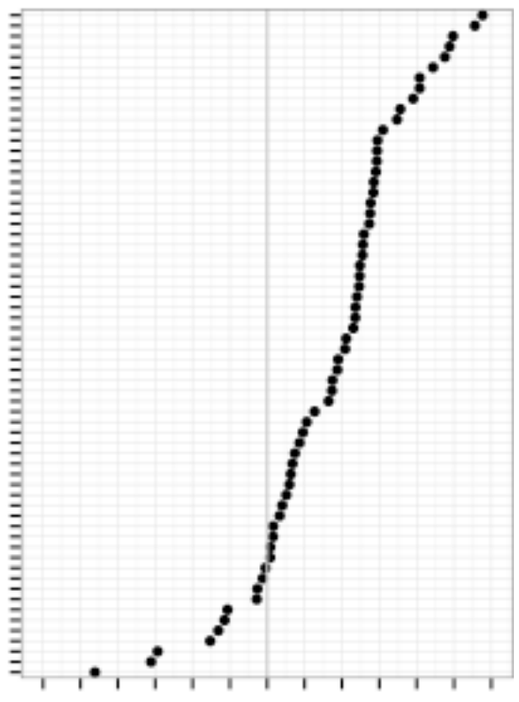
d



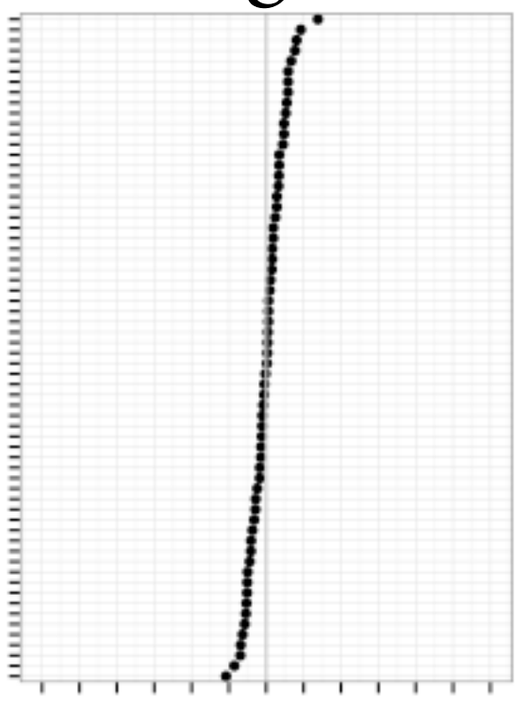
e



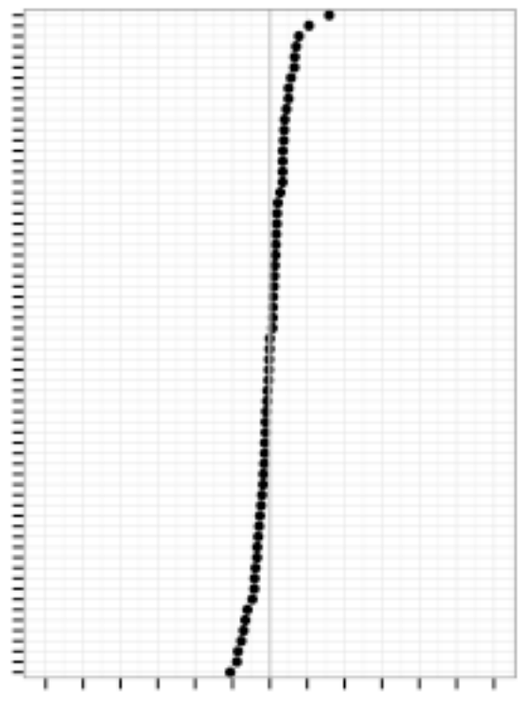
f



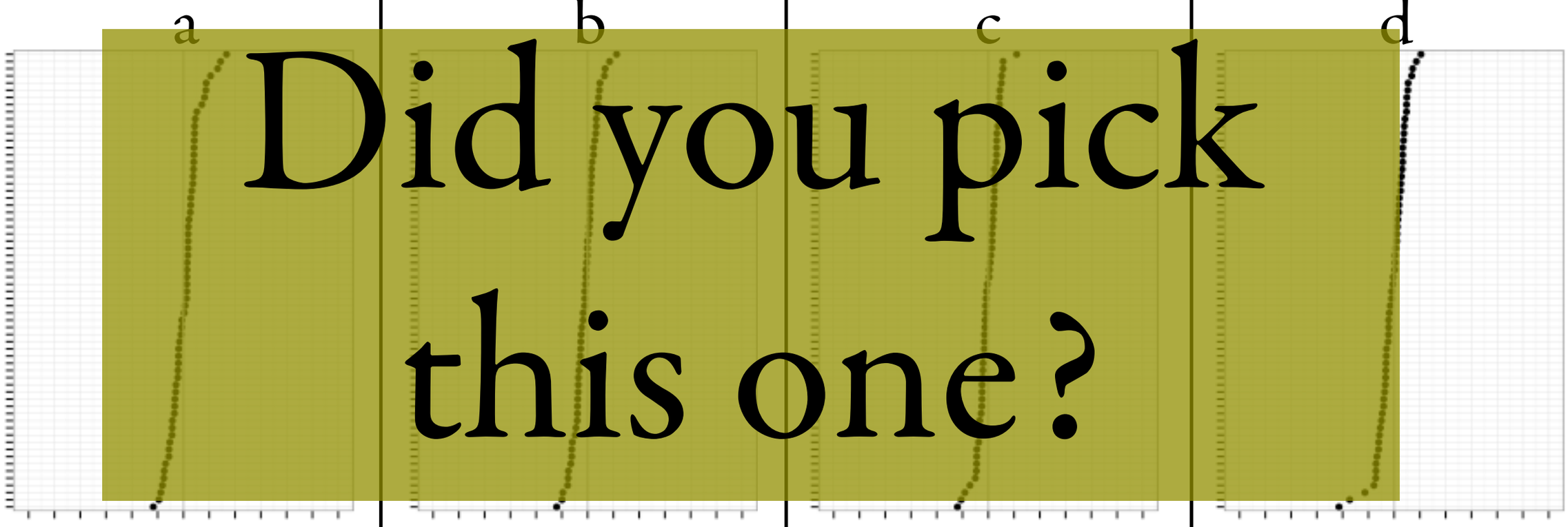
g



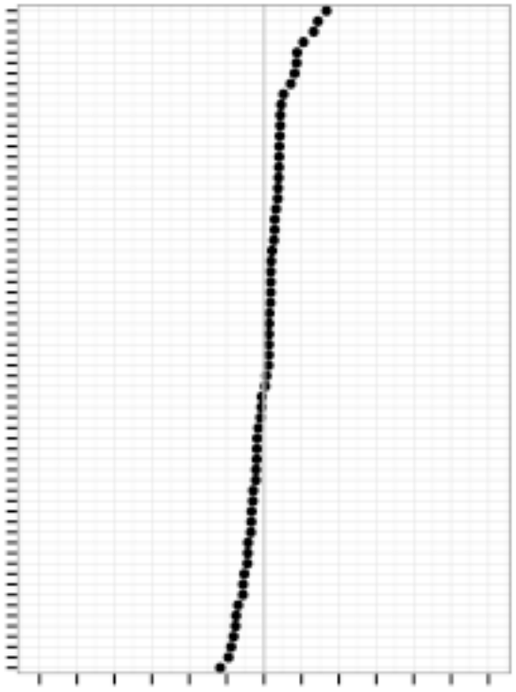
h



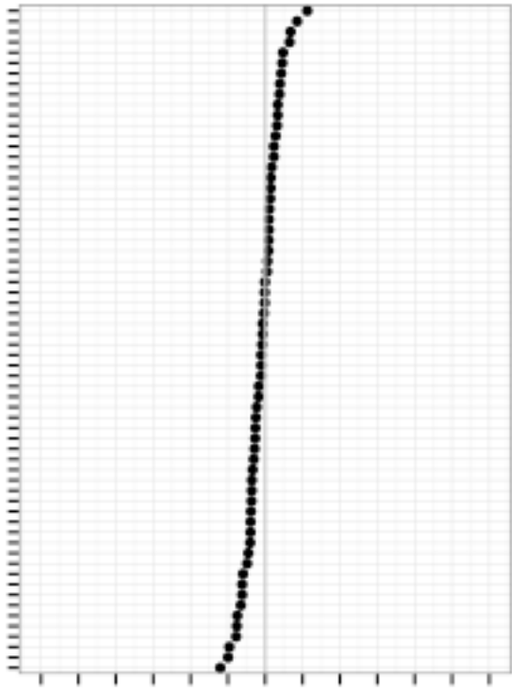
Did you pick this one?



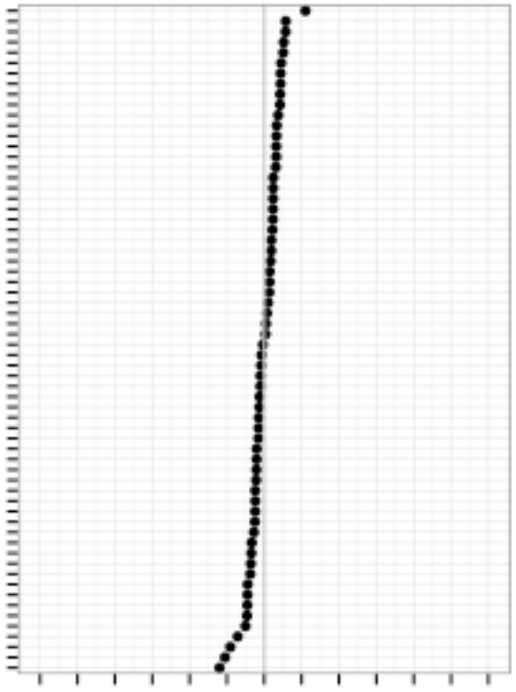
a



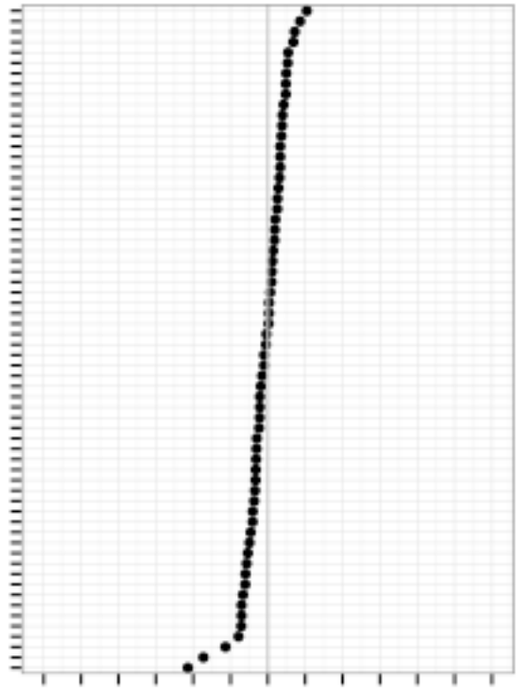
b



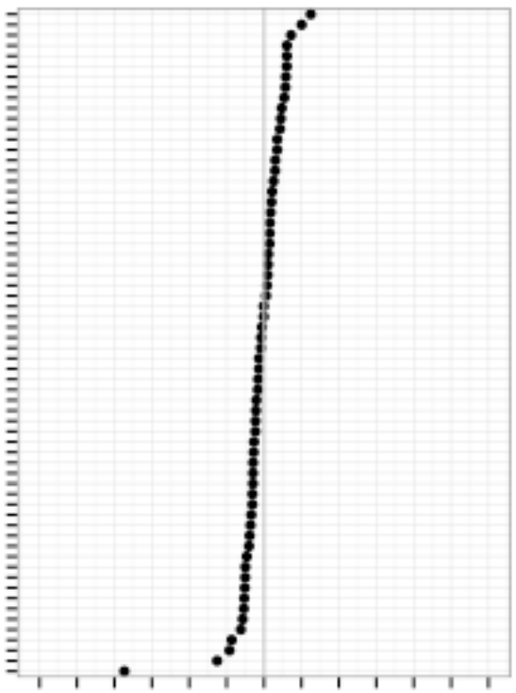
c



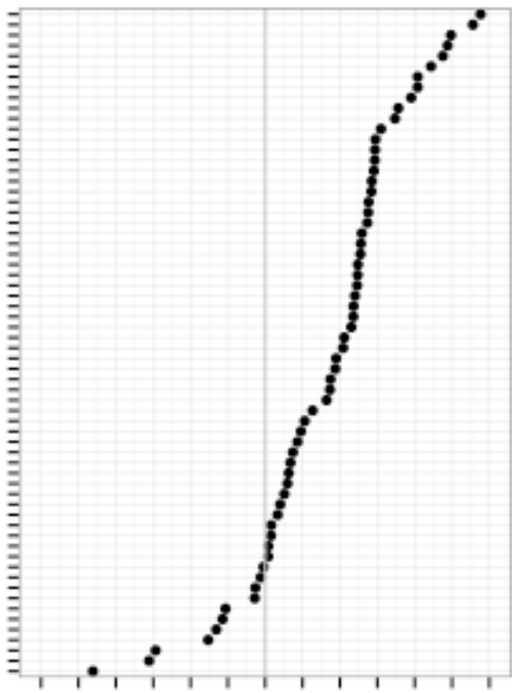
d



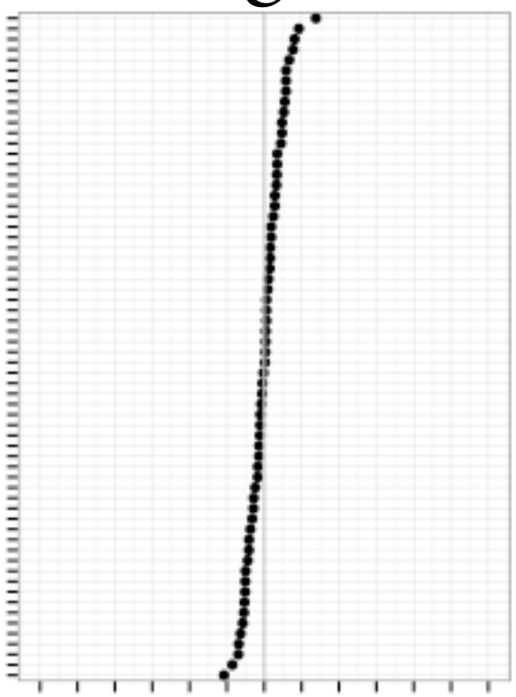
e



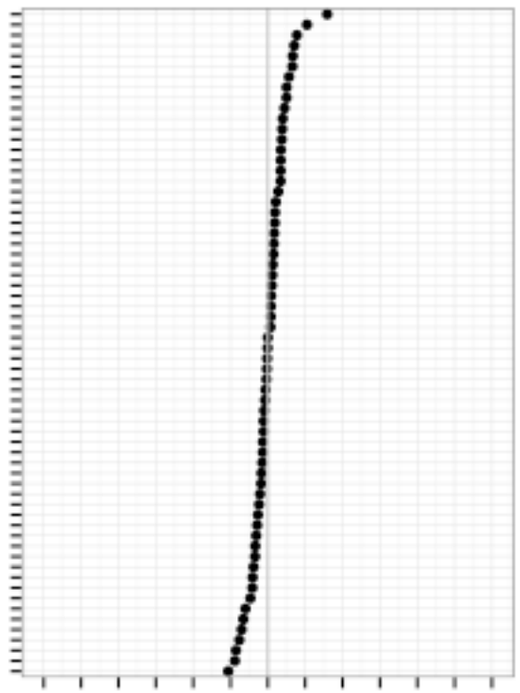
f



g



h



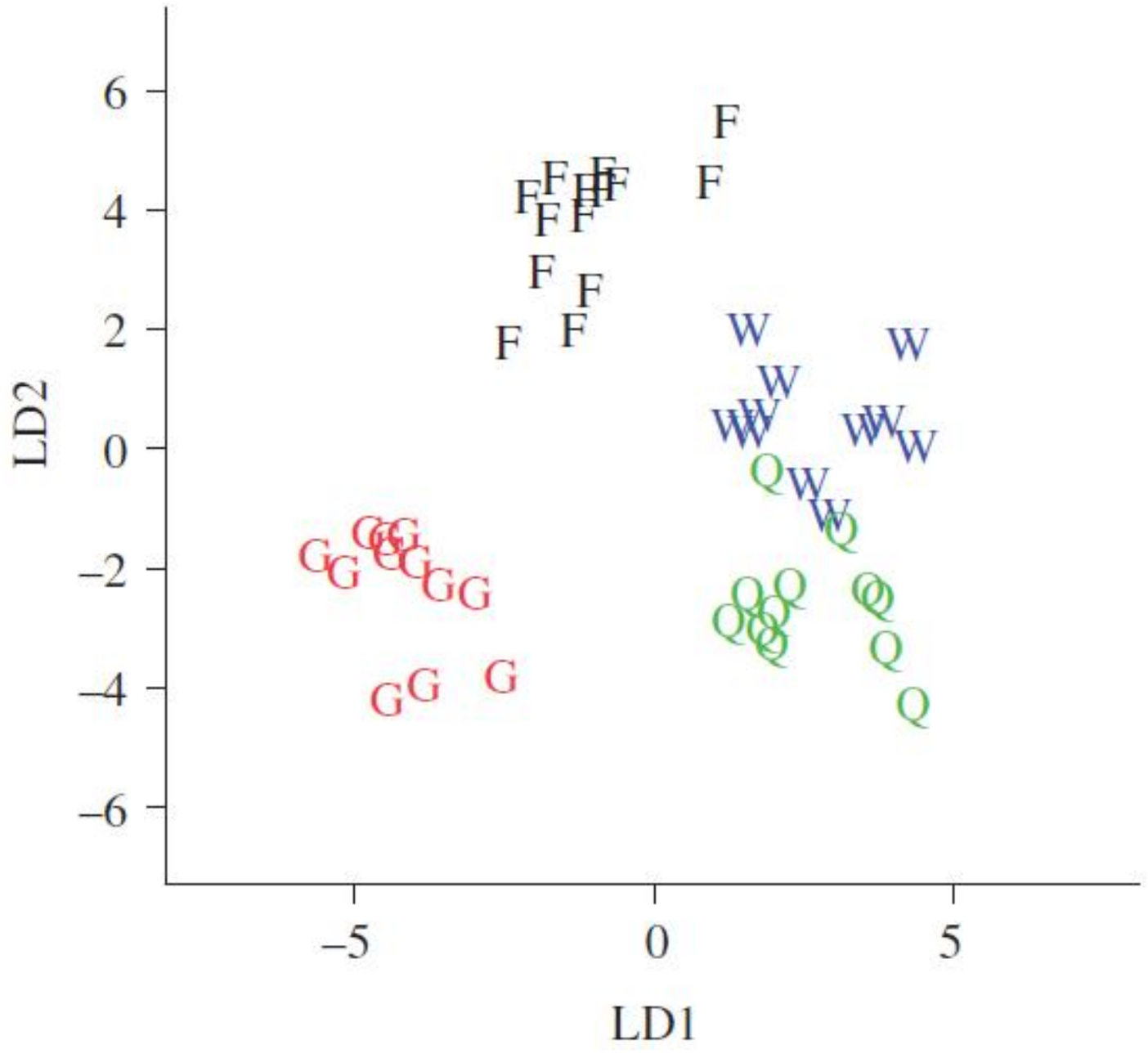
Nulls by permutation

- Hold country fixed (subset by country)
- Permute the gender labels, so that the math scores are randomly assigned to a boy or girl
- Recalculate the difference between the means
- Plot the mean difference by country again

Let's do a real one

- 40 oligos (variables)
- 48 wasps (cases)
- 4 types of wasps
- Best LDA 2D separation of four groups

(Toth et al, 2010)



Really?

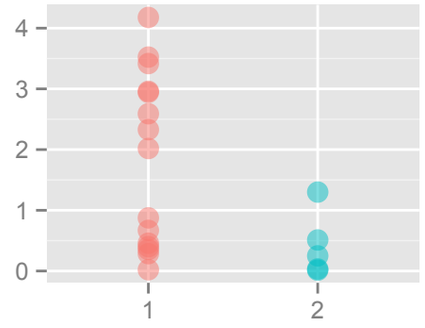
Protocols

- Rorschach: Show many pictures of data with "nothing" happening, pictures from a null distribution
- Lineup: Embed the plot of the data among plots of data generated from the null distribution

Source: Buja et al (2009) RSPT(A)

Protocols

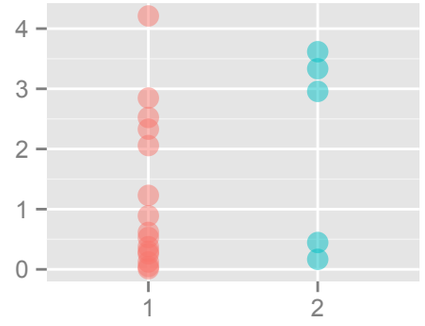
- Rorschach: Show many pictures of data with "nothing" happening, pictures from a null distribution
- Lineup: Embed the plot of the data among plots of data generated from the null distribution



Source: Buja et al (2009) RSPT(A)

Protocols

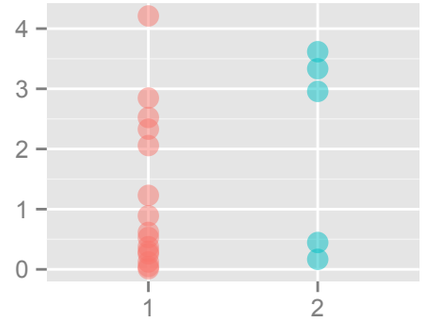
- Rorschach: Show many pictures of data with "nothing" happening, pictures from a null distribution
- Lineup: Embed the plot of the data among plots of data generated from the null distribution



Source: Buja et al (2009) RSPT(A)

Protocols

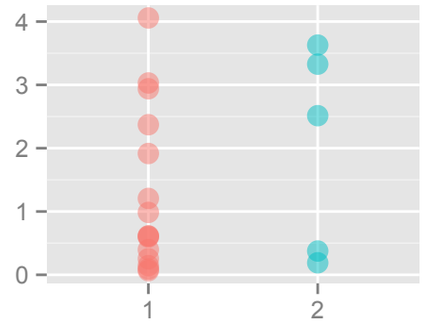
- Rorschach: Show many pictures of data with "nothing" happening, pictures from a null distribution
- Lineup: Embed the plot of the data among plots of data generated from the null distribution



Source: Buja et al (2009) RSPT(A)

Protocols

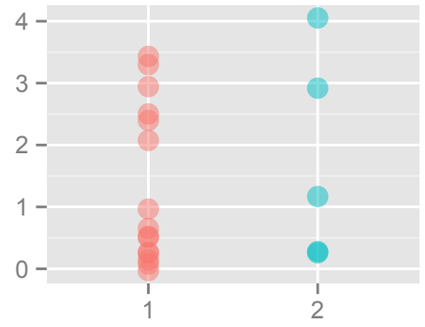
- Rorschach: Show many pictures of data with "nothing" happening, pictures from a null distribution
- Lineup: Embed the plot of the data among plots of data generated from the null distribution



Source: Buja et al (2009) RSPT(A)

Protocols

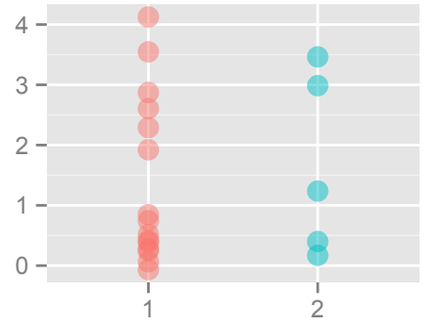
- Rorschach: Show many pictures of data with "nothing" happening, pictures from a null distribution
- Lineup: Embed the plot of the data among plots of data generated from the null distribution



Source: Buja et al (2009) RSPT(A)

Protocols

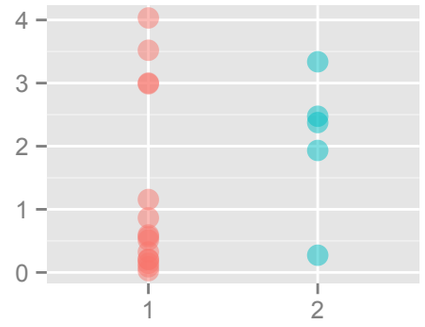
- Rorschach: Show many pictures of data with "nothing" happening, pictures from a null distribution
- Lineup: Embed the plot of the data among plots of data generated from the null distribution



Source: Buja et al (2009) RSPT(A)

Protocols

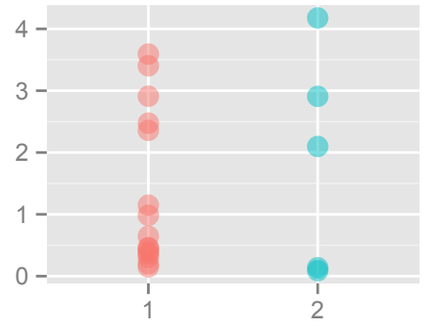
- Rorschach: Show many pictures of data with "nothing" happening, pictures from a null distribution
- Lineup: Embed the plot of the data among plots of data generated from the null distribution



Source: Buja et al (2009) RSPT(A)

Protocols

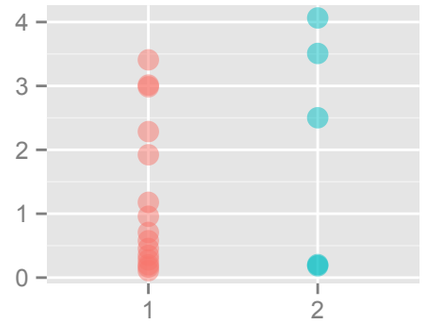
- Rorschach: Show many pictures of data with "nothing" happening, pictures from a null distribution
- Lineup: Embed the plot of the data among plots of data generated from the null distribution



Source: Buja et al (2009) RSPT(A)

Protocols

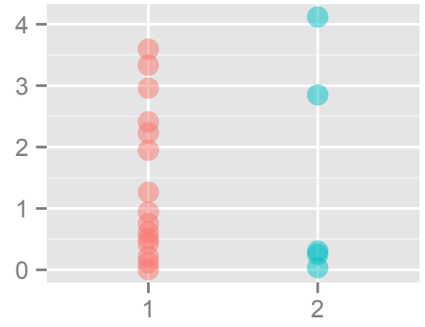
- Rorschach: Show many pictures of data with "nothing" happening, pictures from a null distribution
- Lineup: Embed the plot of the data among plots of data generated from the null distribution



Source: Buja et al (2009) RSPT(A)

Protocols

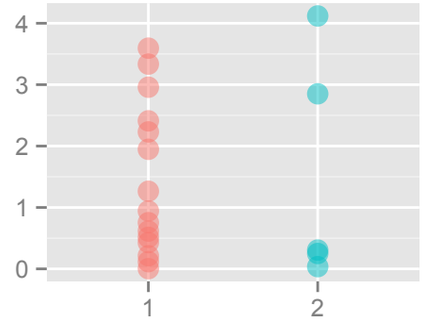
- Rorschach: Show many pictures of data with "nothing" happening, pictures from a null distribution
- Lineup: Embed the plot of the data among plots of data generated from the null distribution

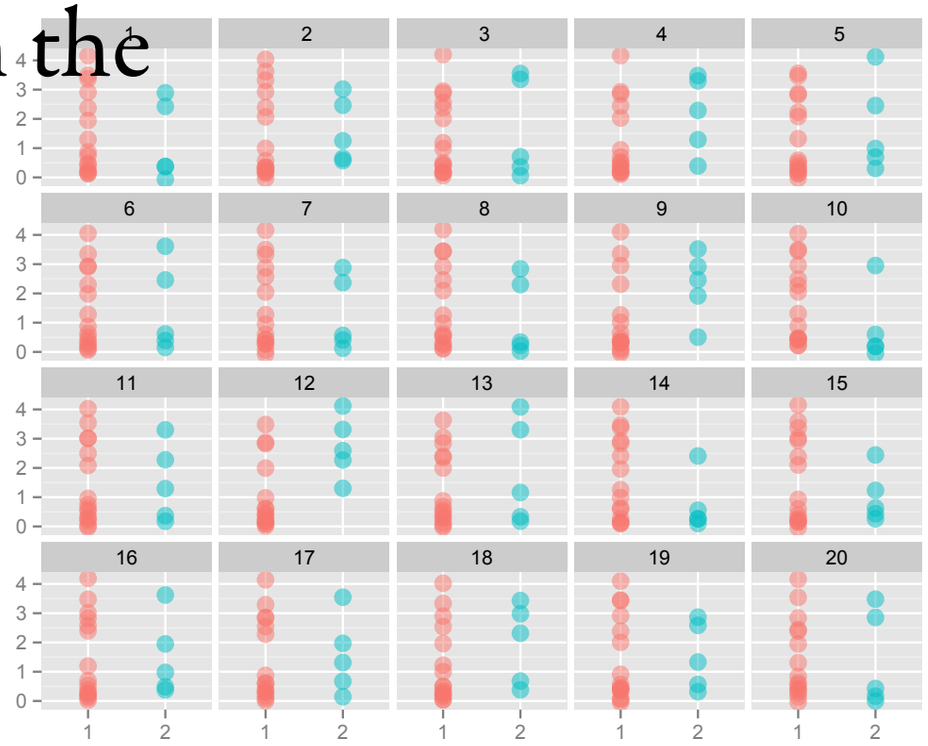


Source: Buja et al (2009) RSPT(A)

Protocols

- Rorschach: Show many pictures of data with "nothing" happening, pictures from a null distribution
- Lineup: Embed the plot of the data among plots of data generated from the null distribution

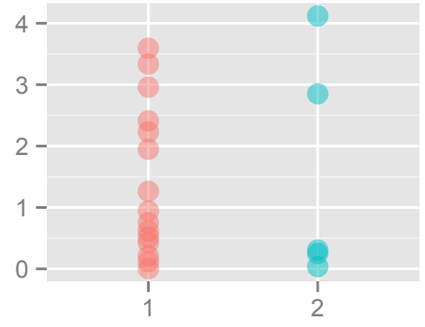


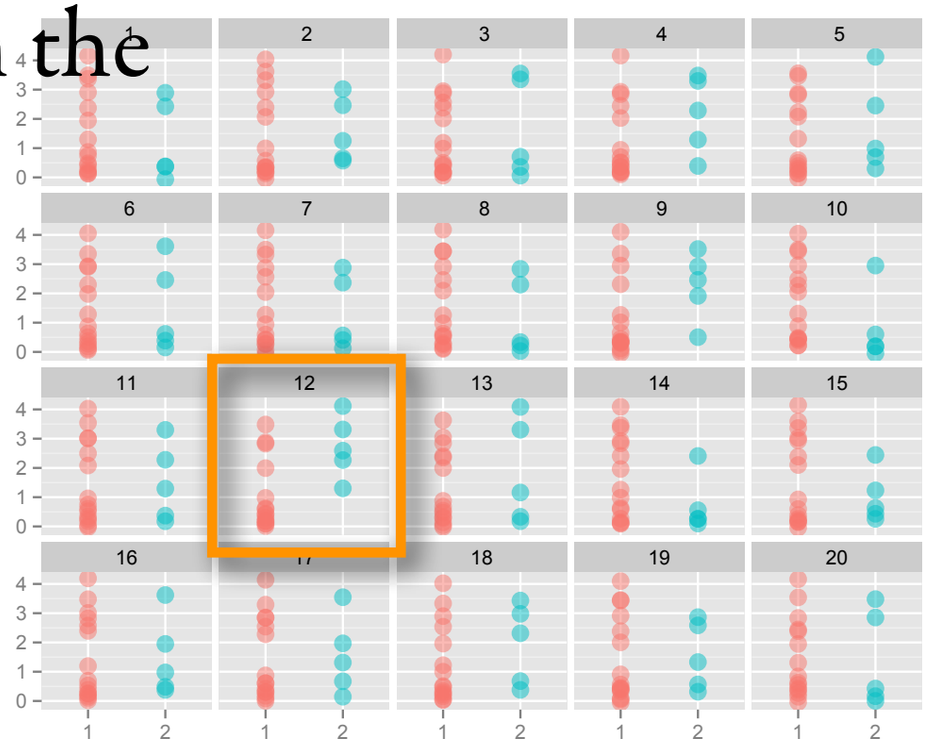


Source: Buja et al (2009) RSPT(A)

Protocols

- Rorschach: Show many pictures of data with "nothing" happening, pictures from a null distribution
- Lineup: Embed the plot of the data among plots of data generated from the null distribution



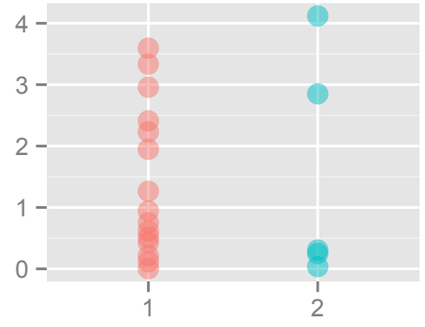


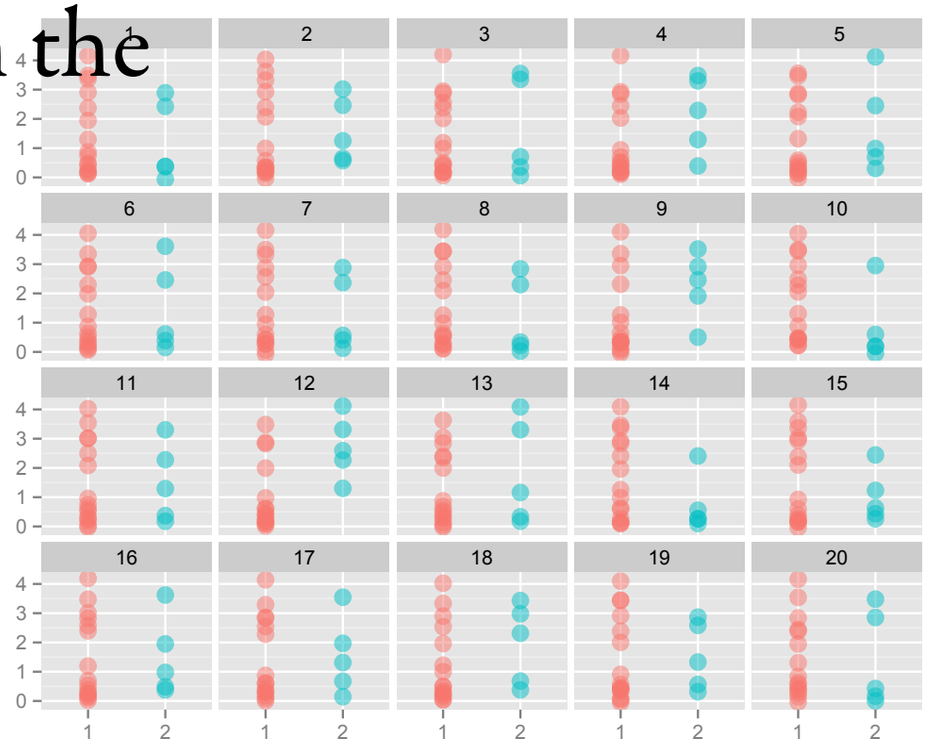
Data plot

Source: Buja et al (2009) RSPT(A)

Protocols

- Rorschach: Show many pictures of data with "nothing" happening, pictures from a null distribution
- Lineup: Embed the plot of the data among plots of data generated from the null distribution

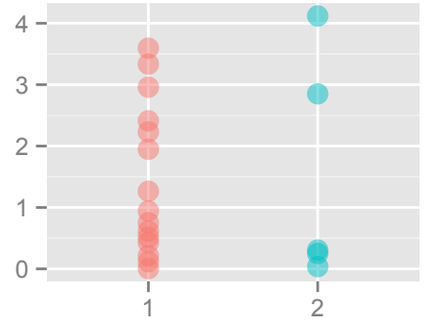


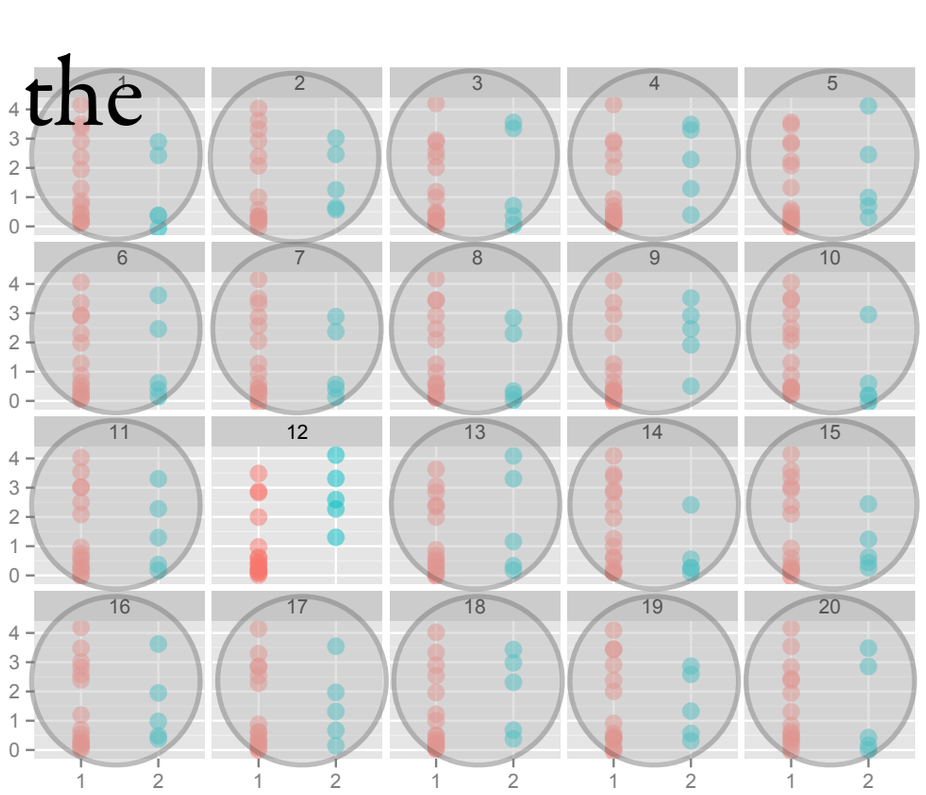


Source: Buja et al (2009) RSPT(A)

Protocols

- Rorschach: Show many pictures of data with "nothing" happening, pictures from a null distribution
- Lineup: Embed the plot of the data among plots of data generated from the null distribution



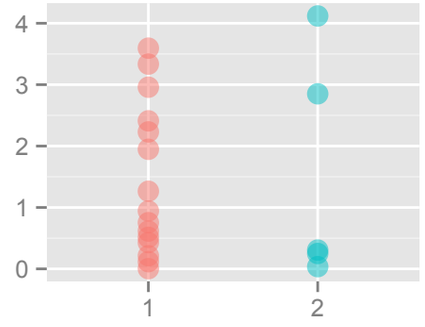


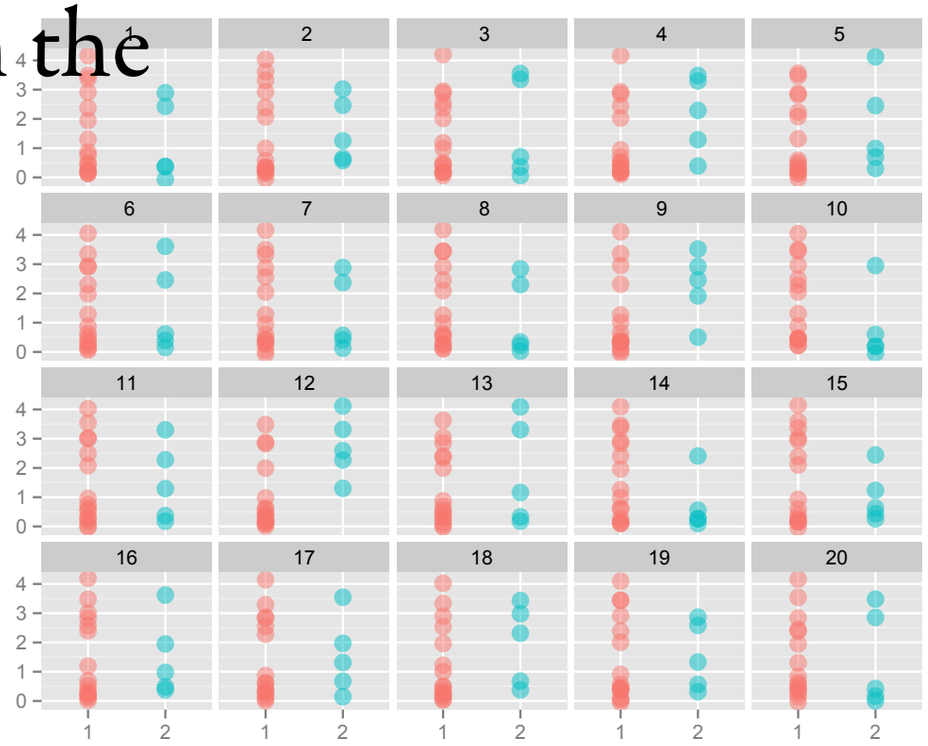
Null plots

Source: Buja et al (2009) RSPT(A)

Protocols

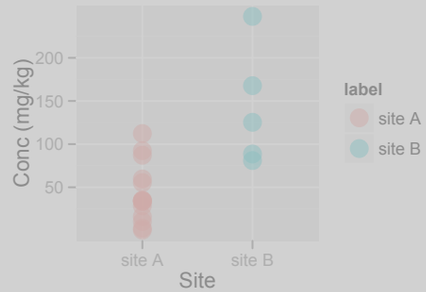
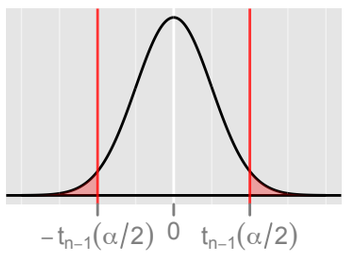
- Rorschach: Show many pictures of data with "nothing" happening, pictures from a null distribution
- Lineup: Embed the plot of the data among plots of data generated from the null distribution



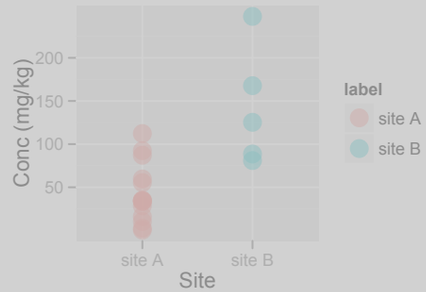
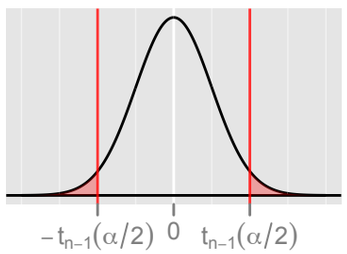


Source: Buja et al (2009) RSPT(A)

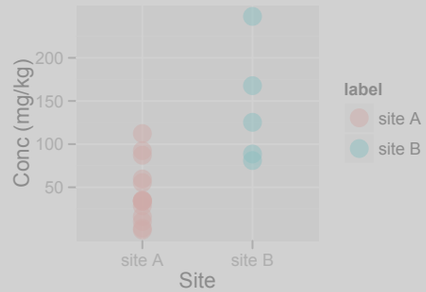
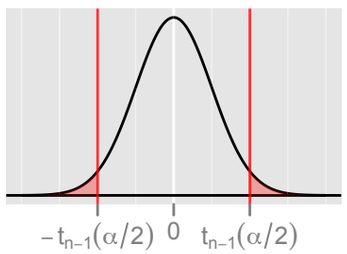
Hypothesis testing

	Mathematical Inference	Visual Inference
Hypothesis	$H_0 : \mu_1 = \mu_2$ vs $H_a : \mu_1 \neq \mu_2$	$H_0 : \mu_1 = \mu_2$ vs $H_a : \mu_1 \neq \mu_2$
	↓	↓
Test Statistic	$T(y) = \frac{\bar{y}_1 - \bar{y}_2}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$	$T(y) =$ 
	↓	↓
Sampling Distribution	$f_{T(y)}(t)$; 	$f_{T(y)}(t)$; 
	↓	↓
Reject H_0 if	observed T is extreme	observed plot is identifiable

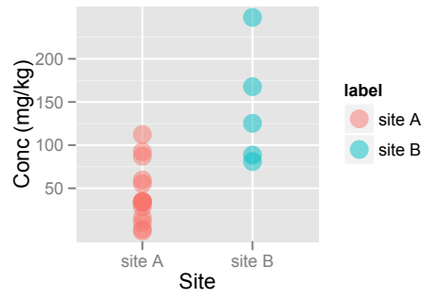
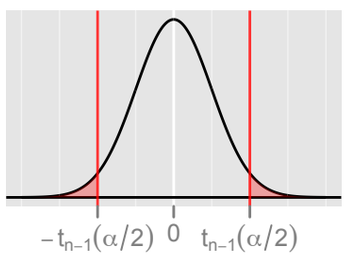
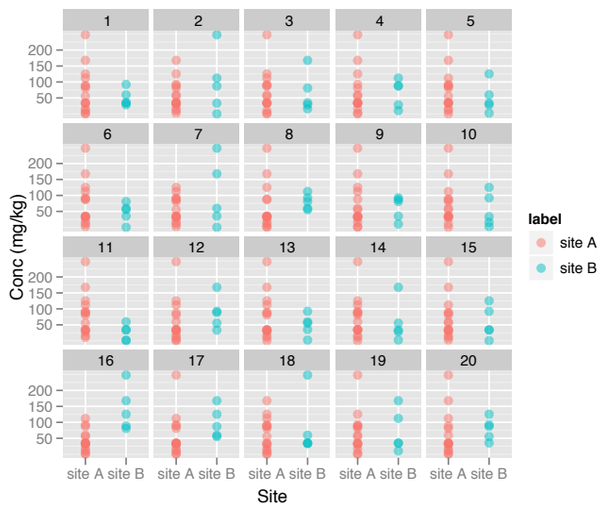
Hypothesis testing

	Mathematical Inference	Visual Inference
Hypothesis	$H_0 : \mu_1 = \mu_2$ vs $H_a : \mu_1 \neq \mu_2$	$H_0 : \mu_1 = \mu_2$ vs $H_a : \mu_1 \neq \mu_2$
	↓	↓
Test Statistic	$T(y) = \frac{\bar{y}_1 - \bar{y}_2}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$	$T(y) =$ 
	↓	↓
Sampling Distribution	$f_{T(y)}(t)$; 	$f_{T(y)}(t)$; 
	↓	↓
Reject H_0 if	observed T is extreme	observed plot is identifiable

Hypothesis testing

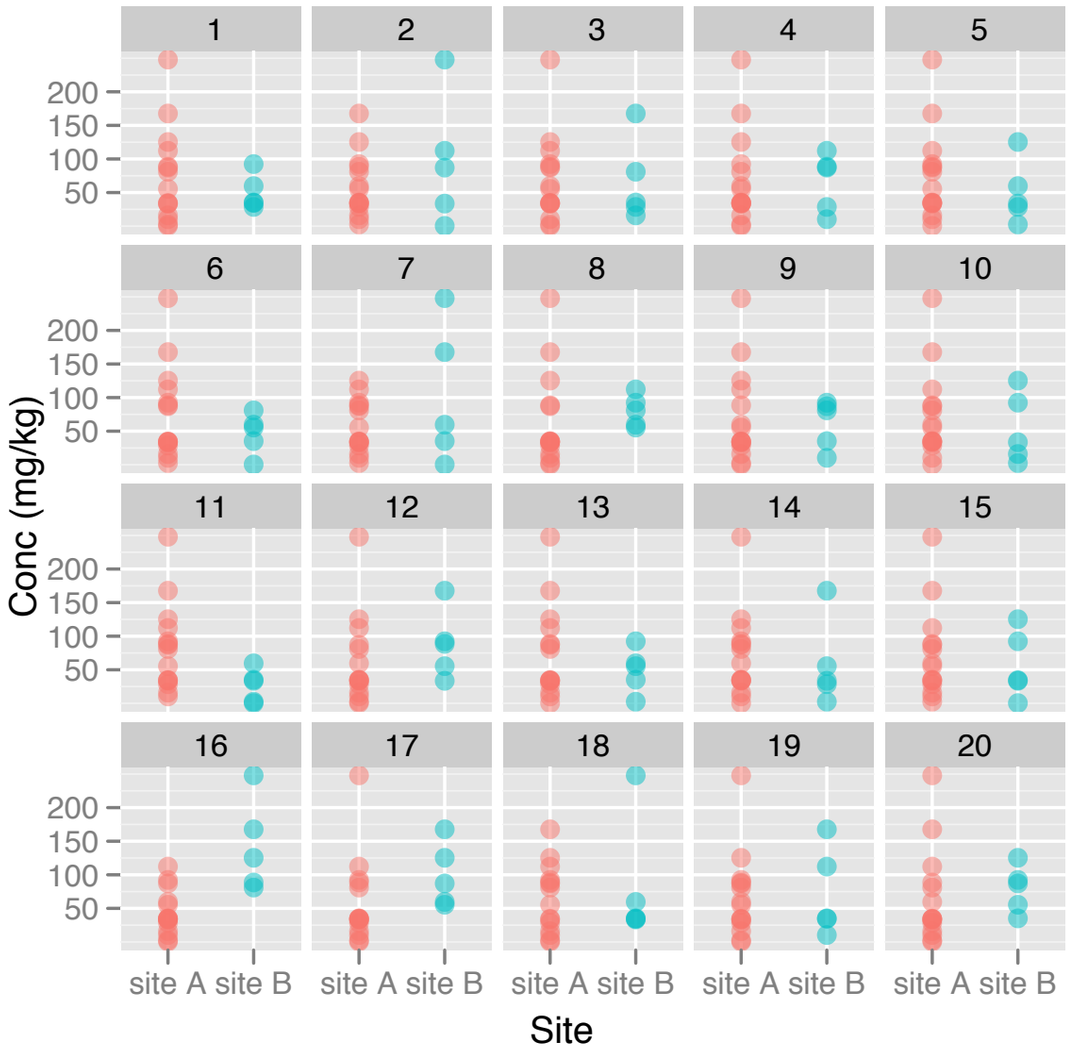
	Mathematical Inference	Visual Inference
Hypothesis	$H_0 : \mu_1 = \mu_2$ vs $H_a : \mu_1 \neq \mu_2$	$H_0 : \mu_1 = \mu_2$ vs $H_a : \mu_1 \neq \mu_2$
	↓	↓
Test Statistic	$T(y) = \frac{\bar{y}_1 - \bar{y}_2}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$	$T(y) =$ 
	↓	↓
Sampling Distribution	$f_{T(y)}(t)$; 	$f_{T(y)}(t)$; 
	↓	↓
Reject H_0 if	observed T is extreme	observed plot is identifiable

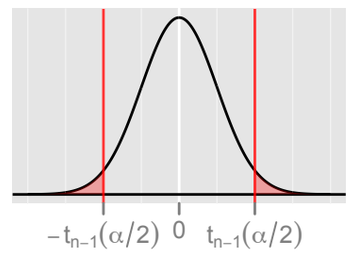
Hypothesis testing

	Mathematical Inference	Visual Inference
Hypothesis	$H_0 : \mu_1 = \mu_2$ vs $H_a : \mu_1 \neq \mu_2$	$H_0 : \mu_1 = \mu_2$ vs $H_a : \mu_1 \neq \mu_2$
Test Statistic	$T(y) = \frac{\bar{y}_1 - \bar{y}_2}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$	$T(y) =$ 
Sampling Distribution	$f_{T(y)}(t);$ 	$f_{T(y)}(t);$ 
Reject H_0 if	observed T is extreme	observed plot is identifiable

Consideration ONE

Sampling distribution comparison is against a finite

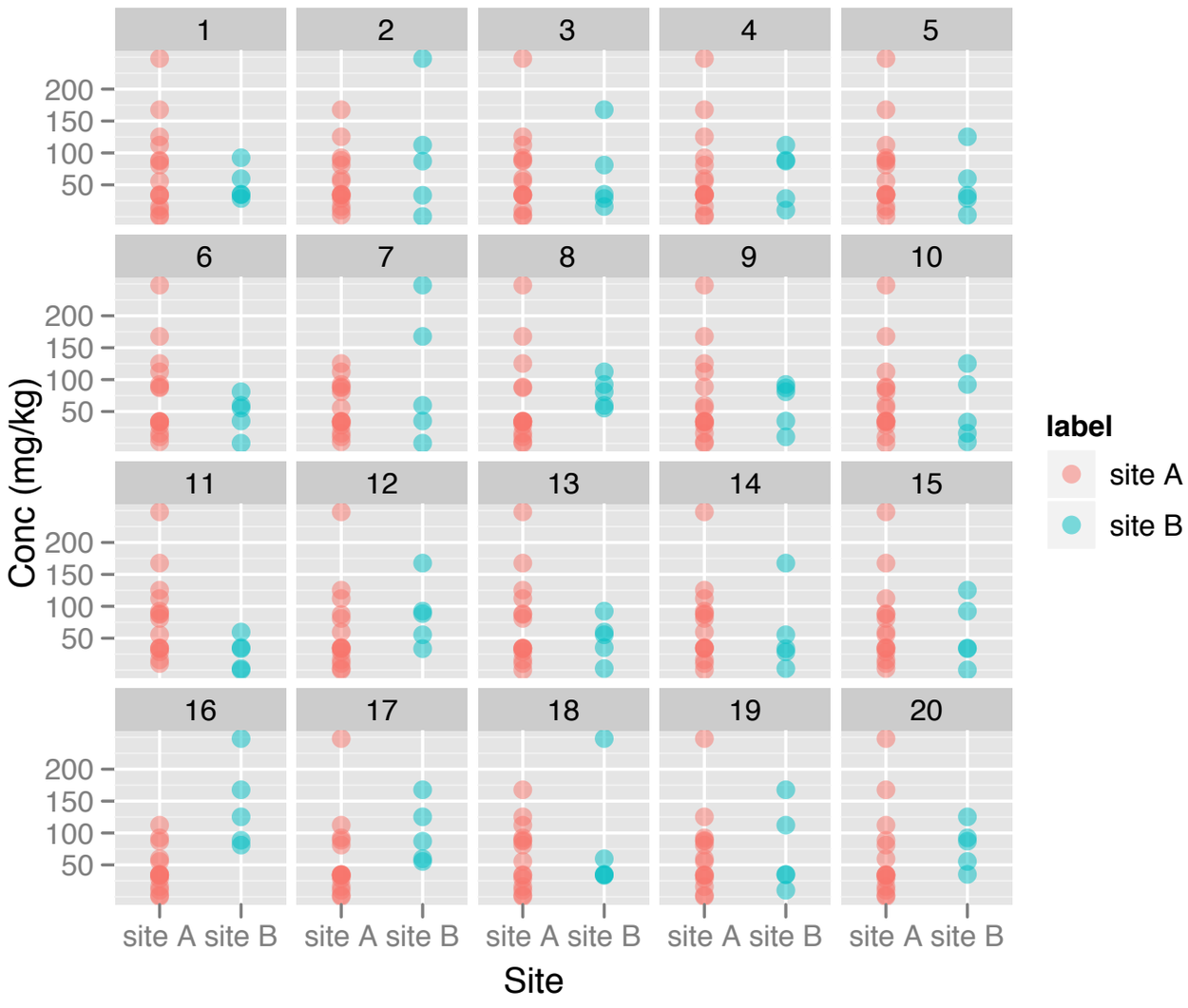


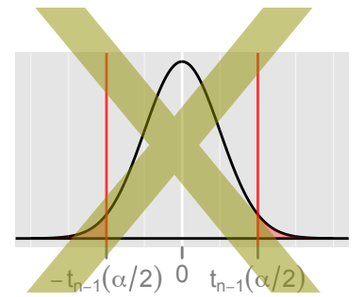


Source: Roy Chowdhury (2014)

Consideration ONE

Sampling distribution comparison is against a finite

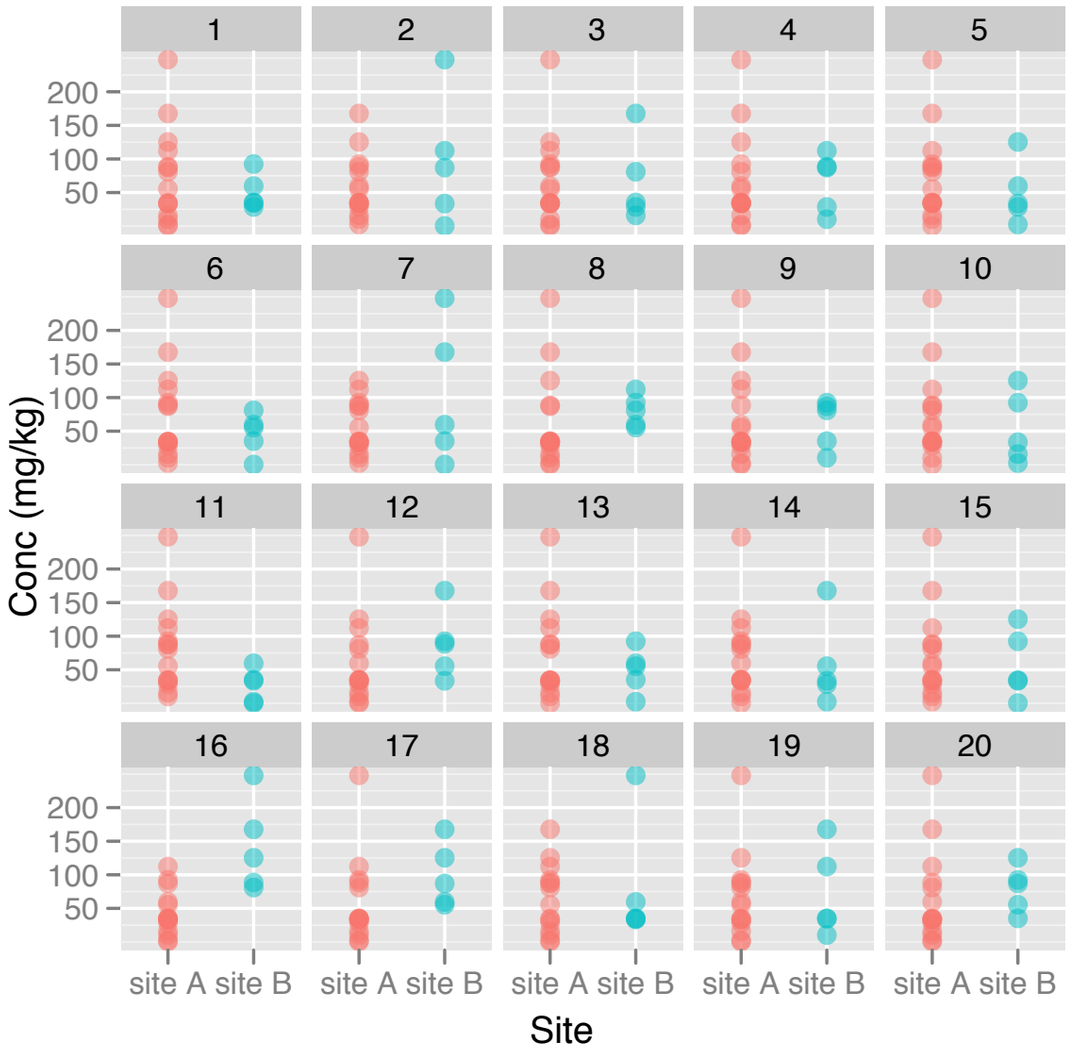


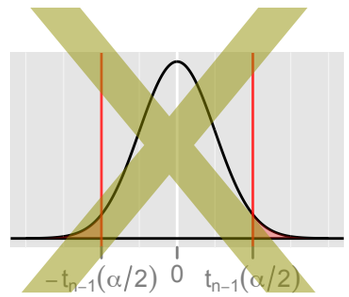


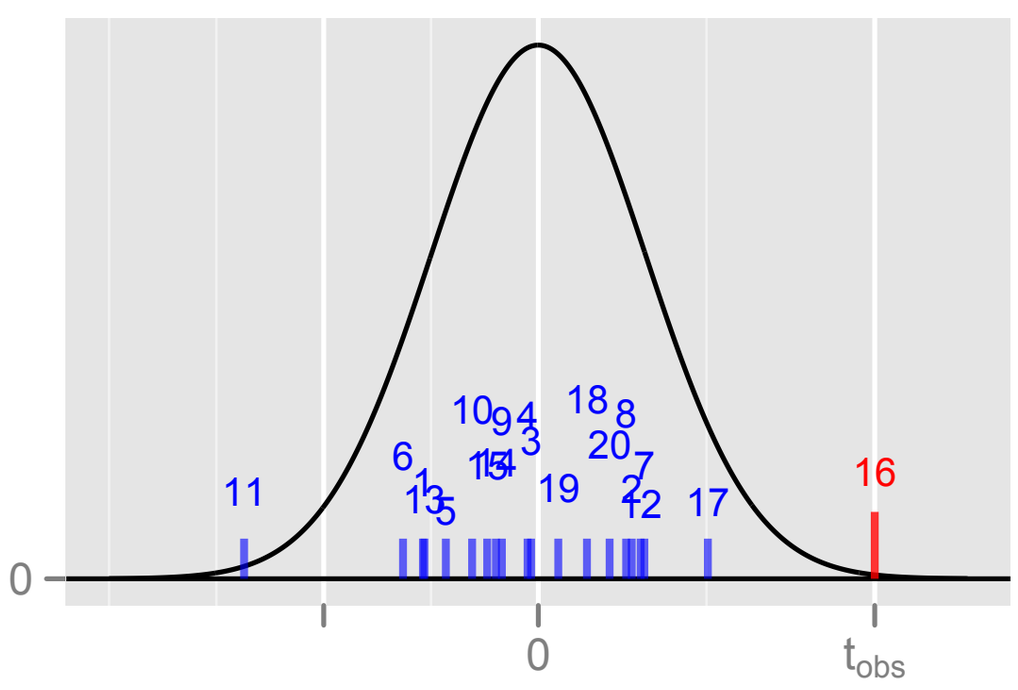
Source: Roy Chowdhury (2014)

Consideration ONE

Sampling distribution comparison is against a finite



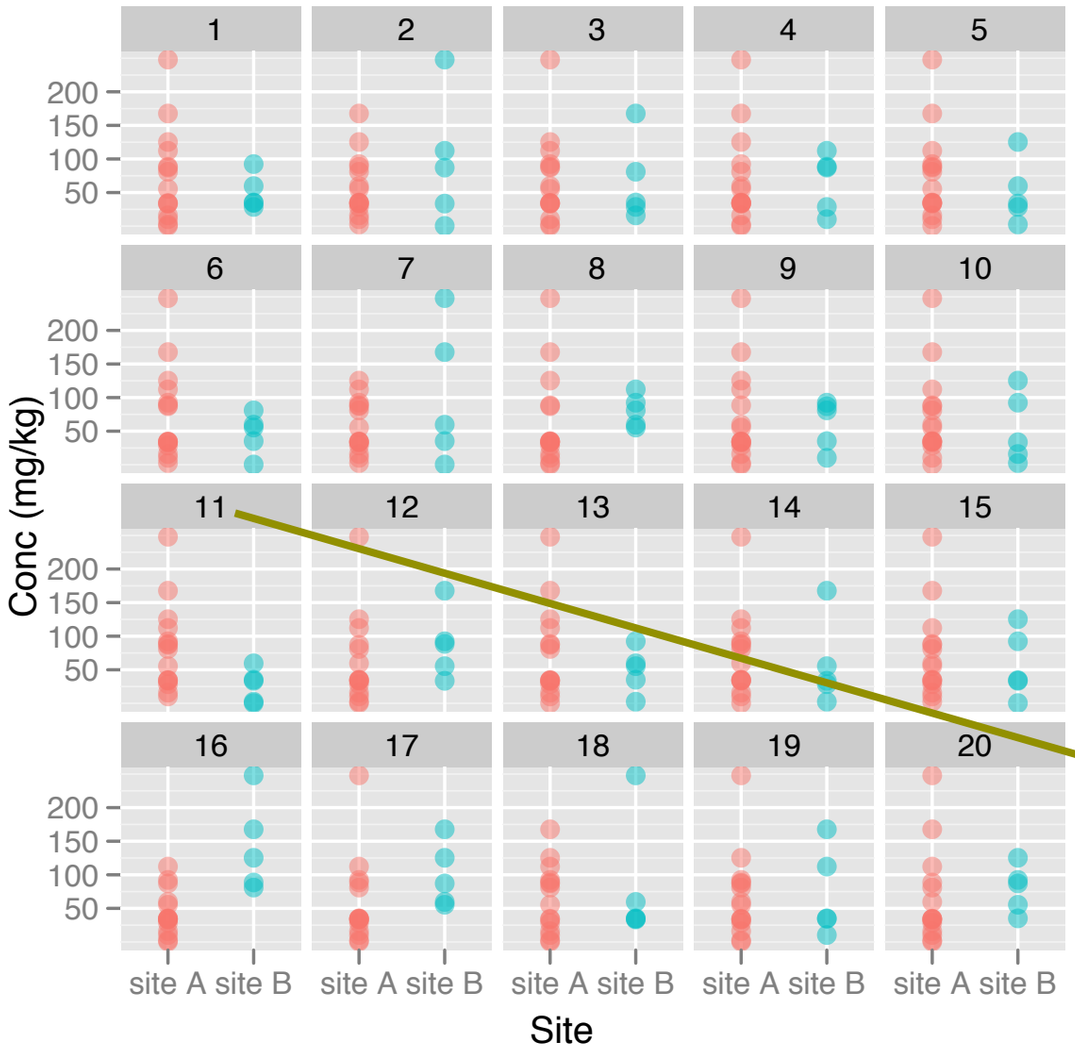


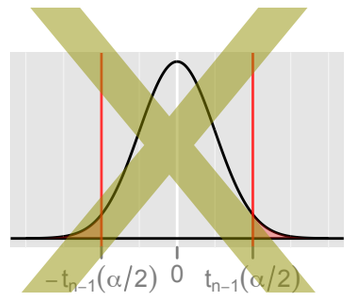


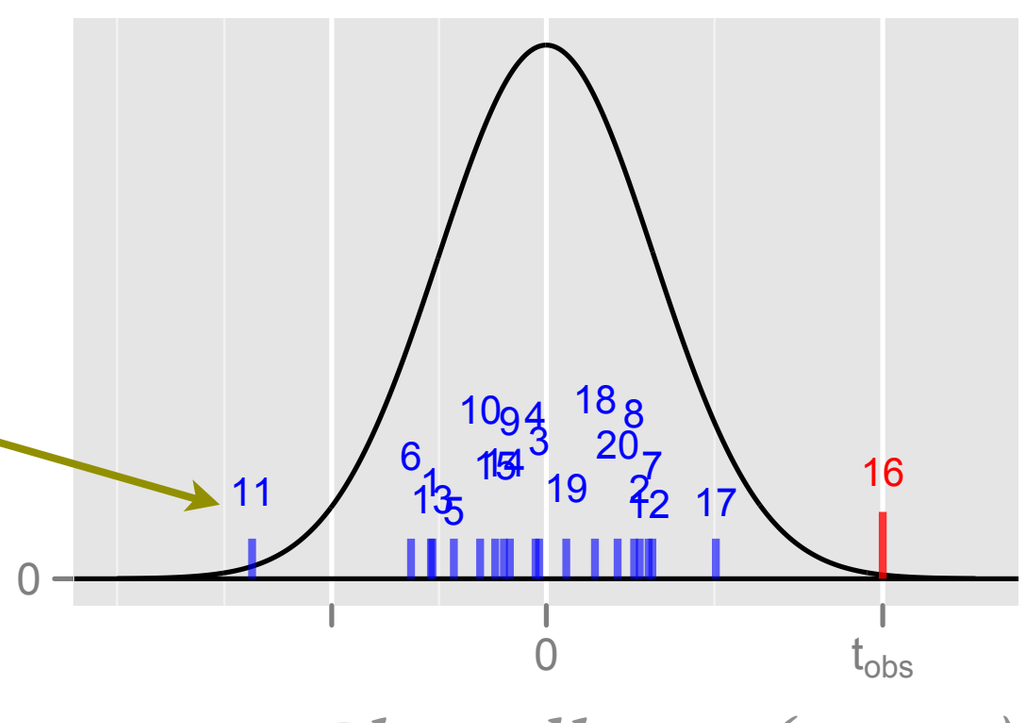
Source: Roy Chowdhury (2014)

Consideration ONE

Sampling distribution comparison is against a finite



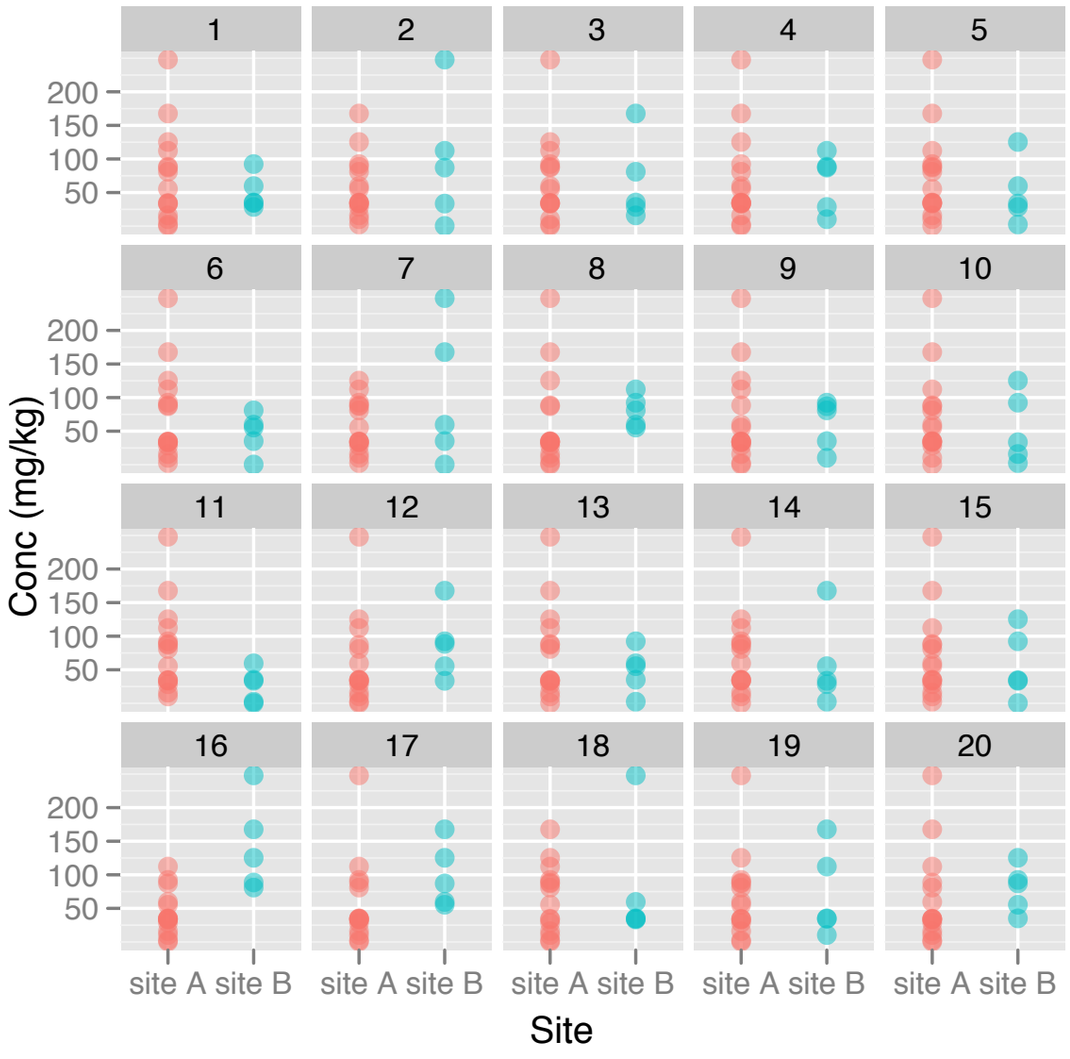


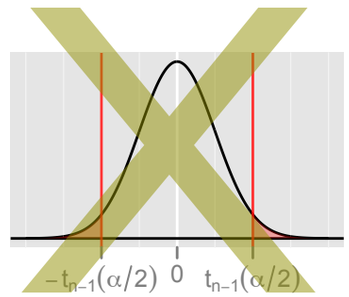


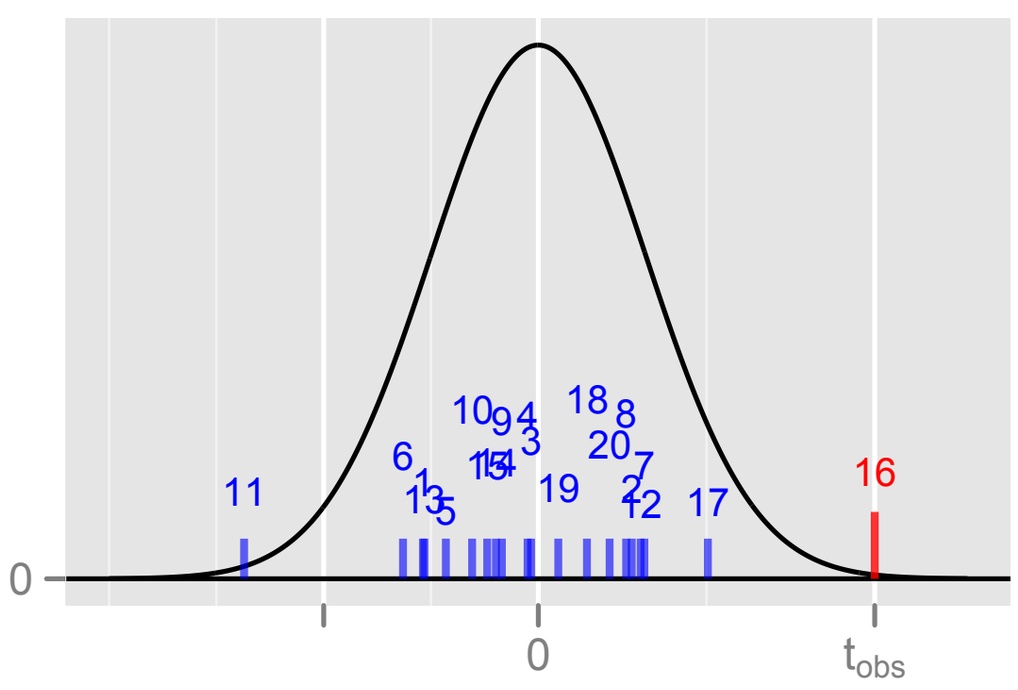
Source: Roy Chowdhury (2014)

Consideration ONE

Sampling distribution comparison is against a finite



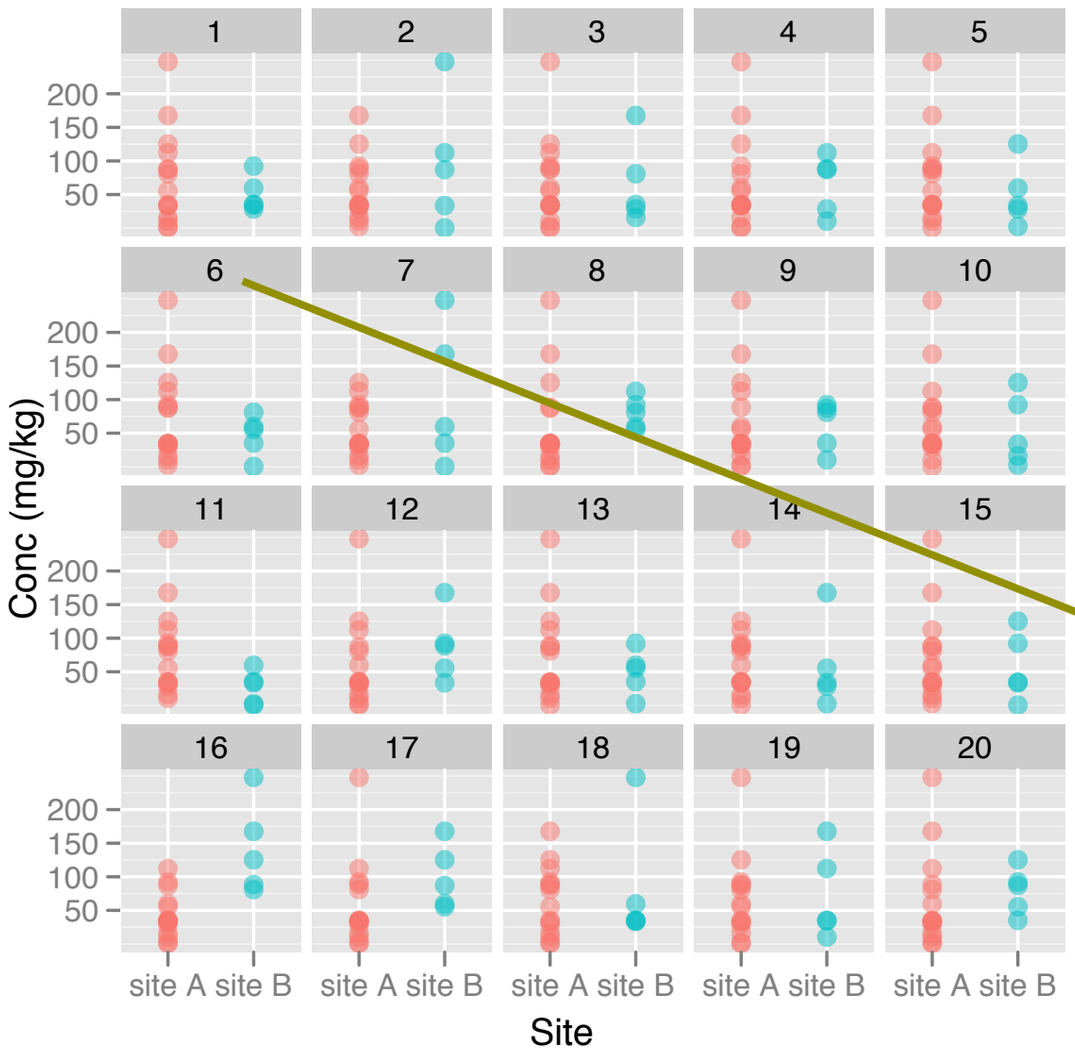


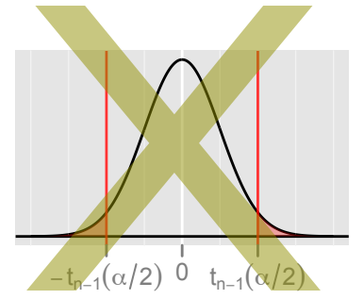


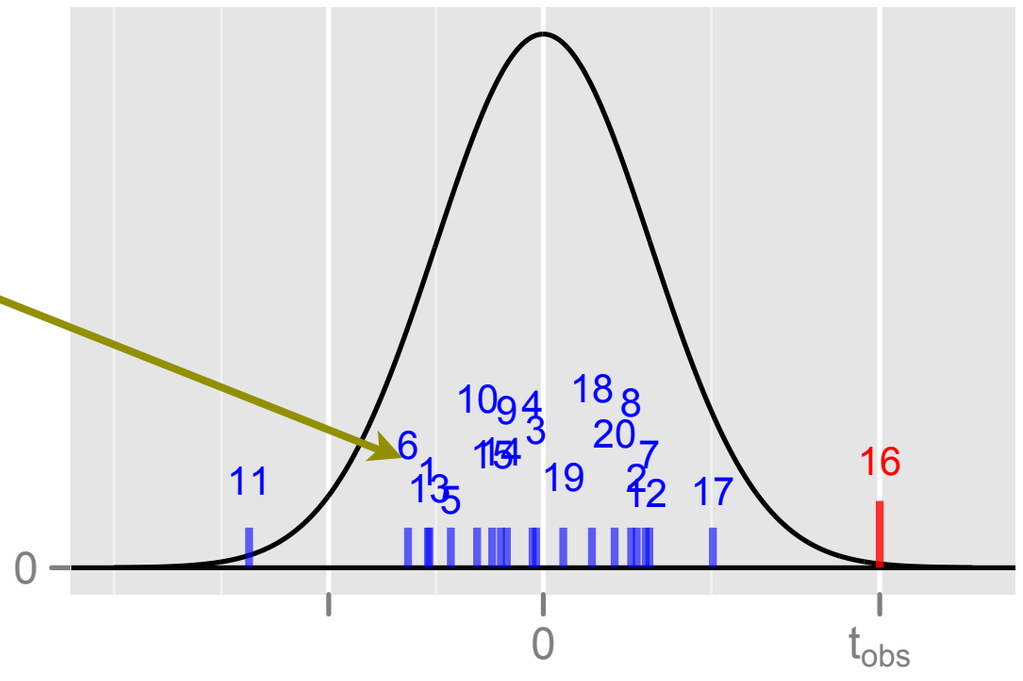
Source: Roy Chowdhury (2014)

Consideration ONE

Sampling distribution comparison is against a finite



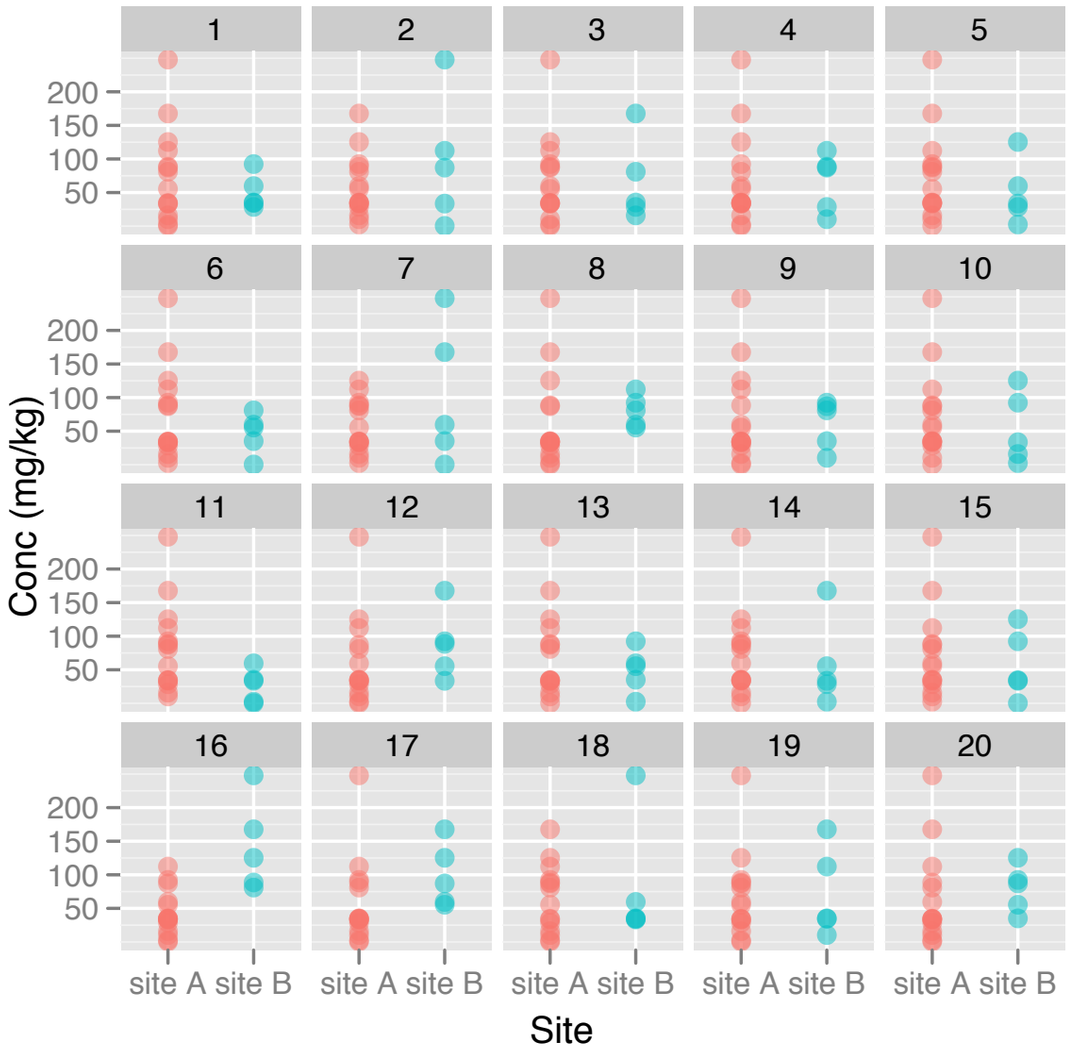


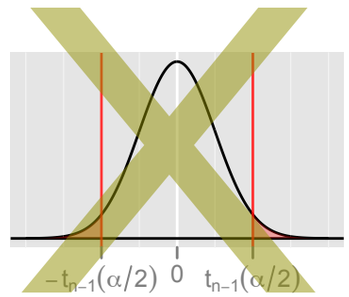


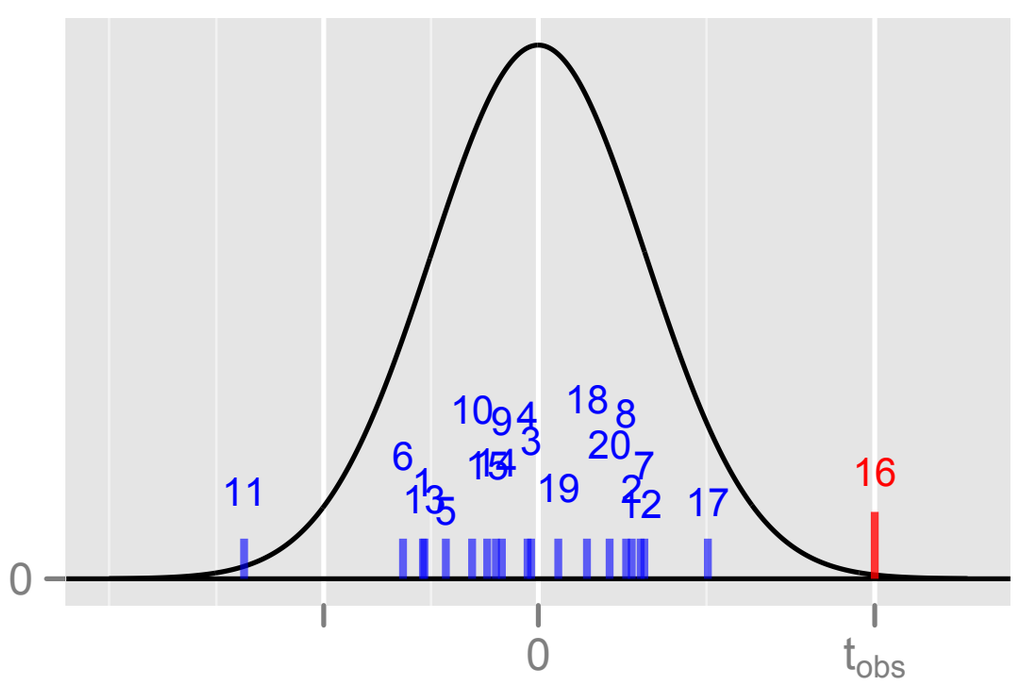
Source: Roy Chowdhury (2014)

Consideration ONE

Sampling distribution comparison is against a finite



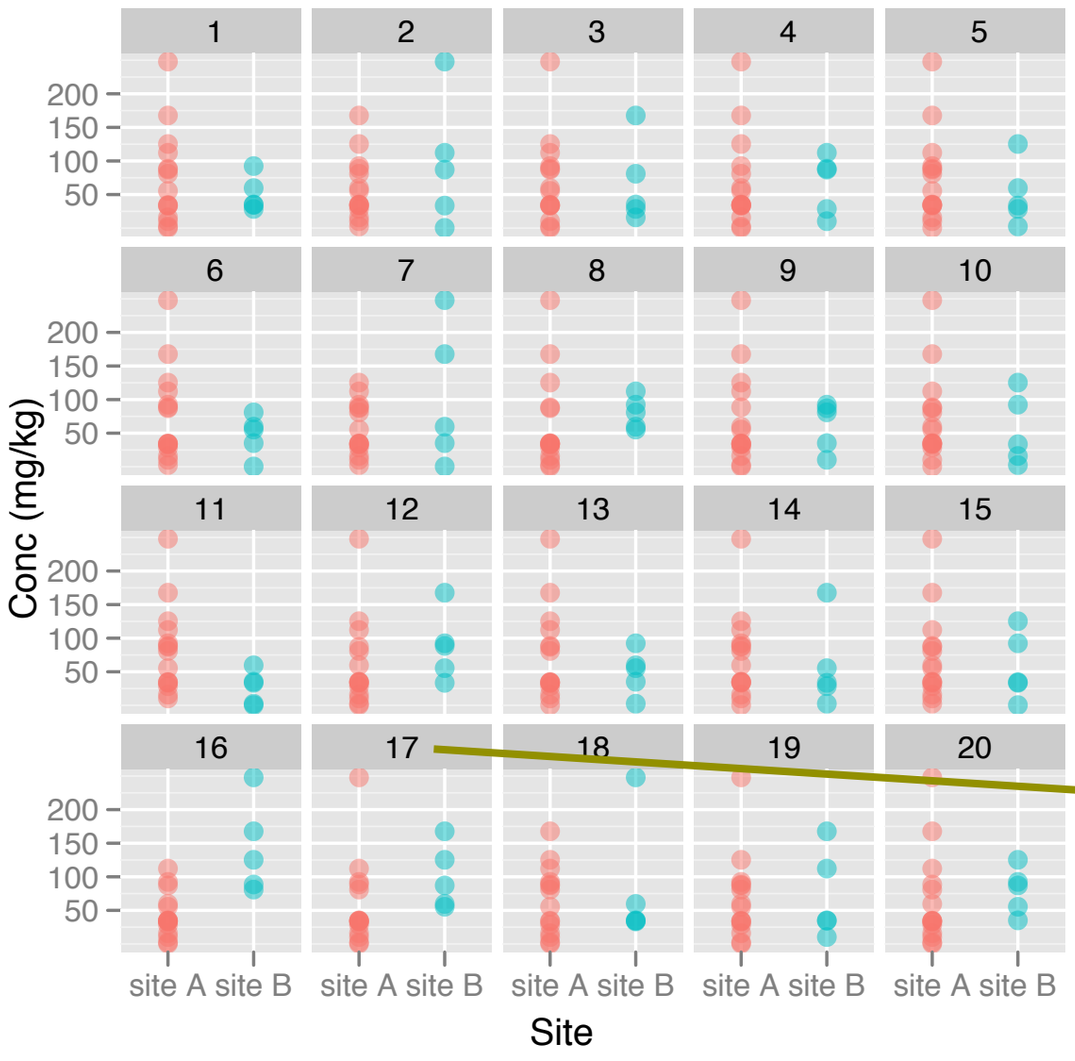


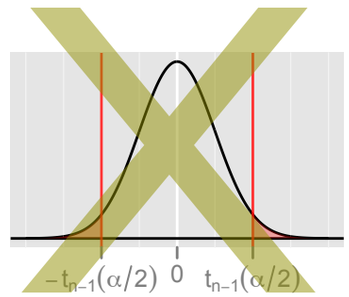


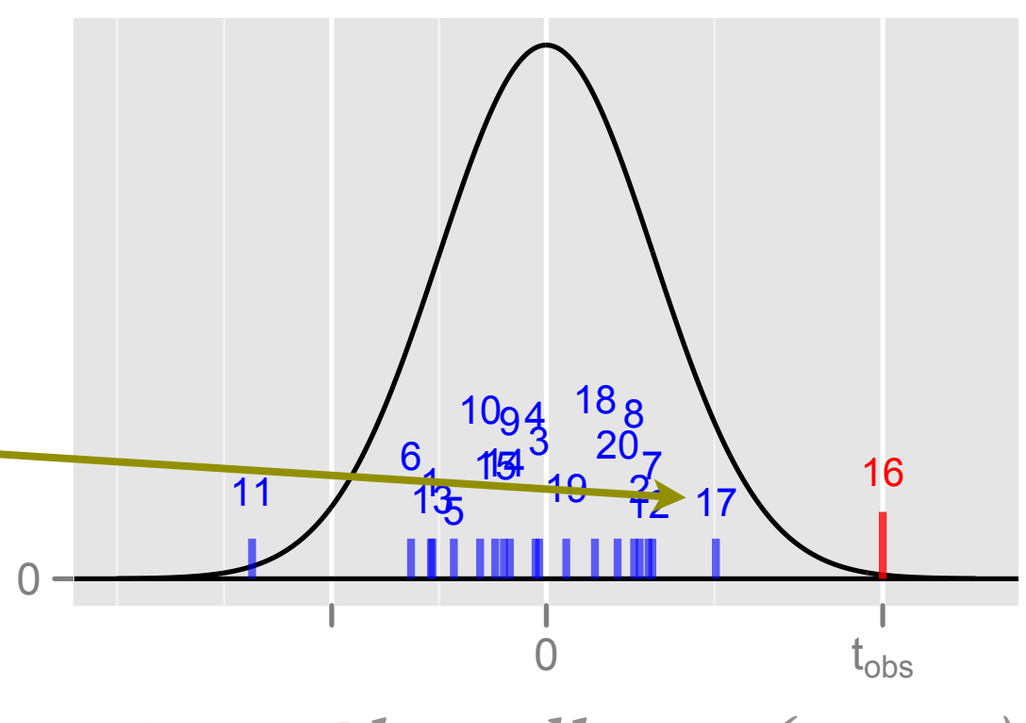
Source: Roy Chowdhury (2014)

Consideration ONE

Sampling distribution comparison is against a finite



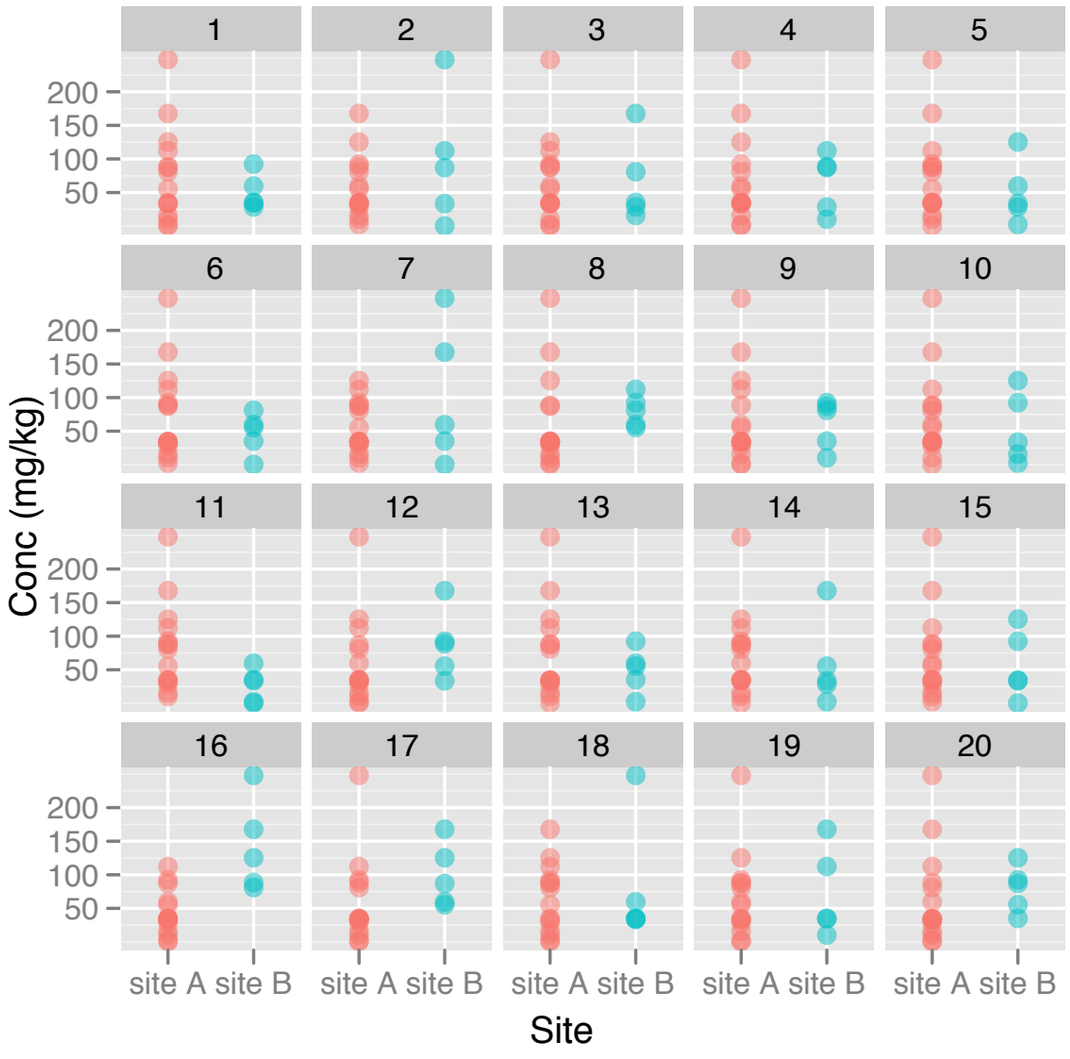


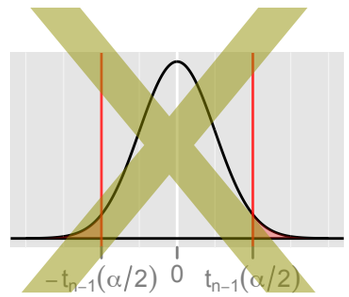


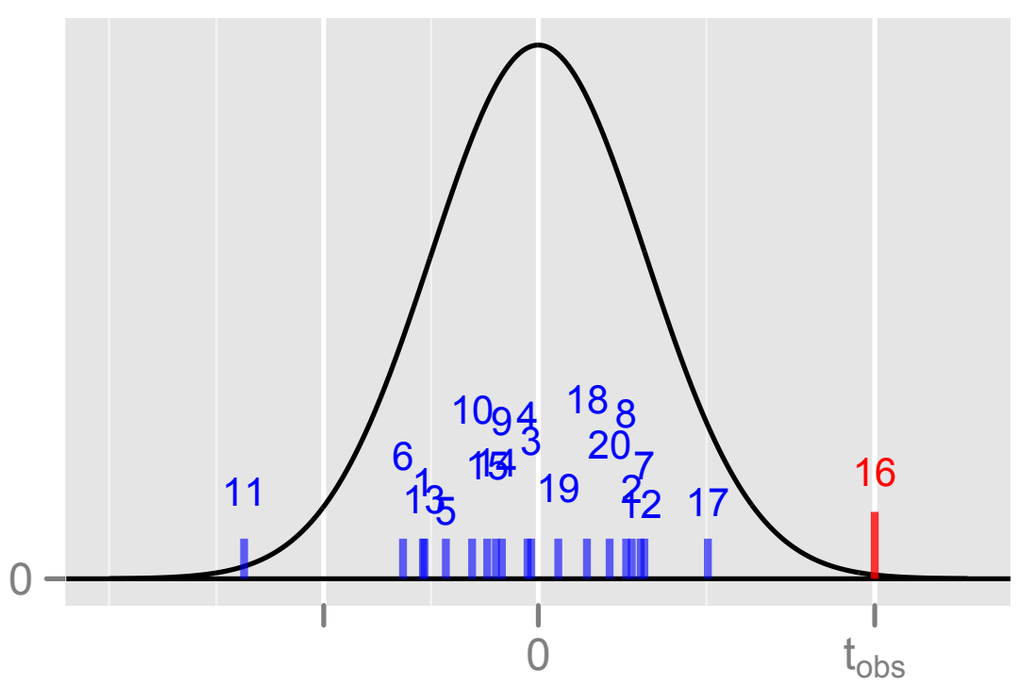
Source: Roy Chowdhury (2014)

Consideration ONE

Sampling distribution comparison is against a finite



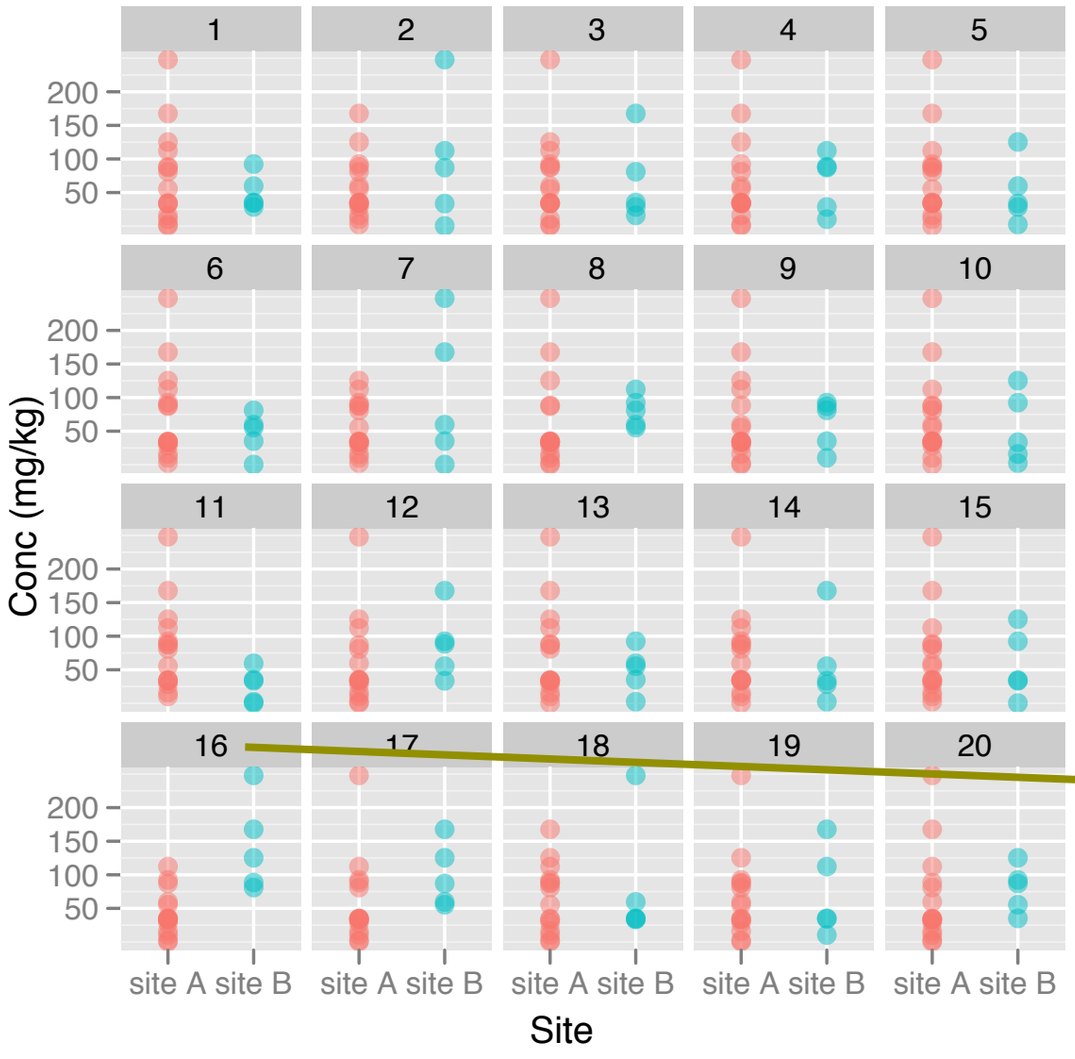


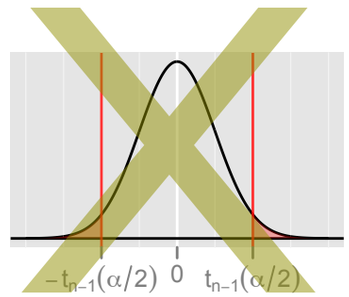


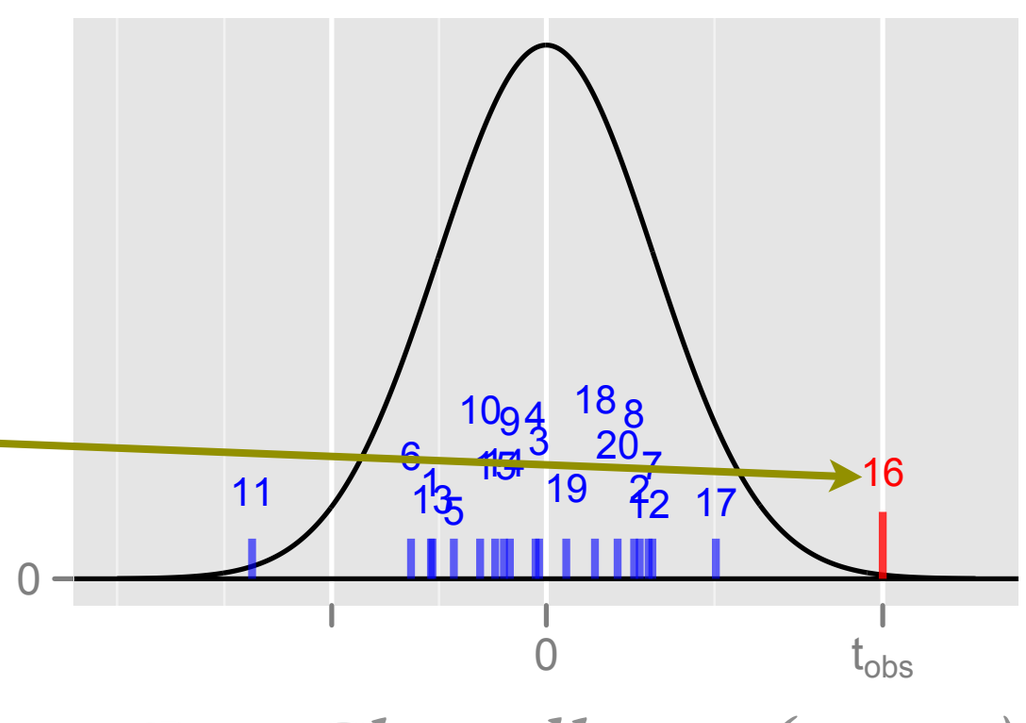
Source: Roy Chowdhury (2014)

Consideration ONE

Sampling distribution comparison is against a finite



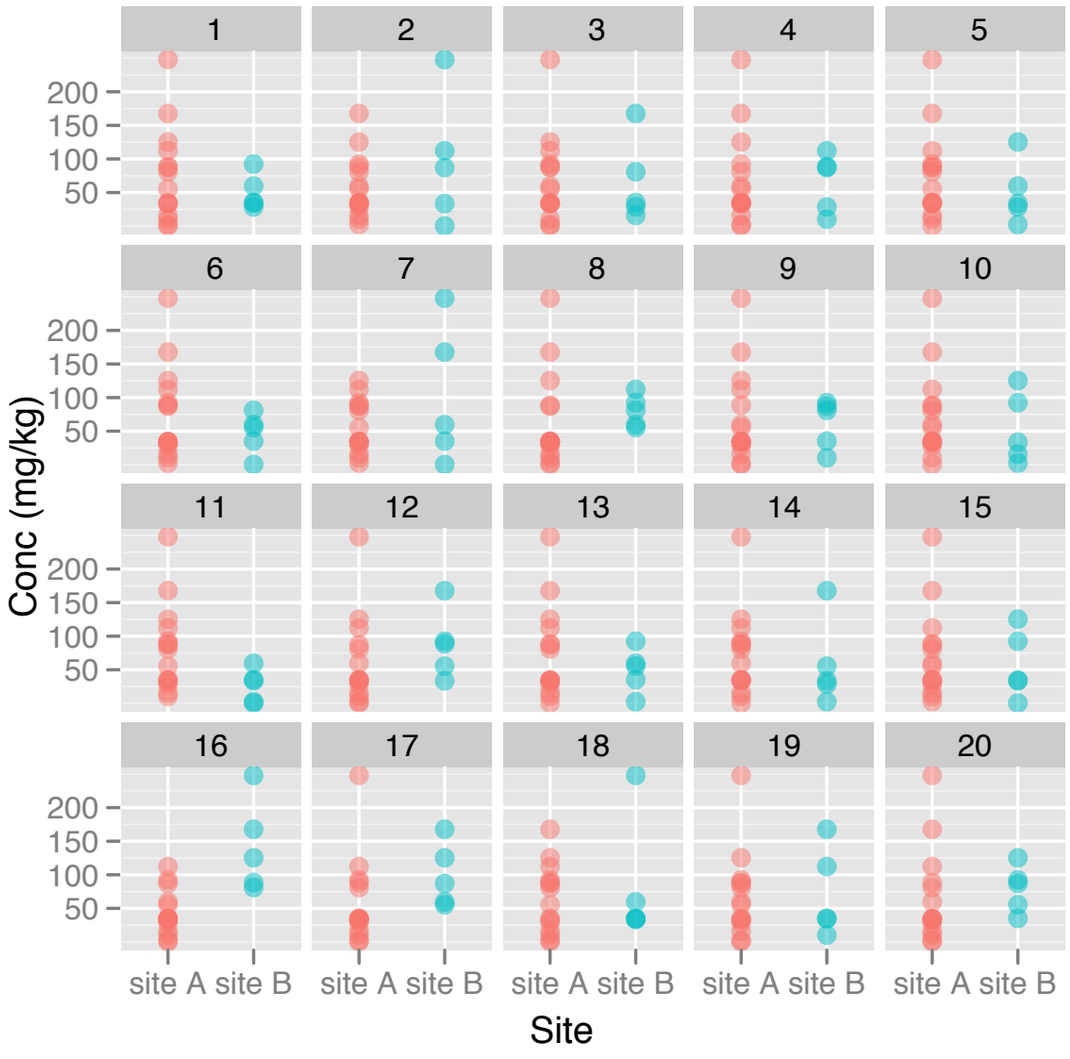


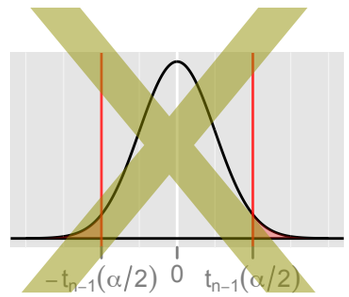


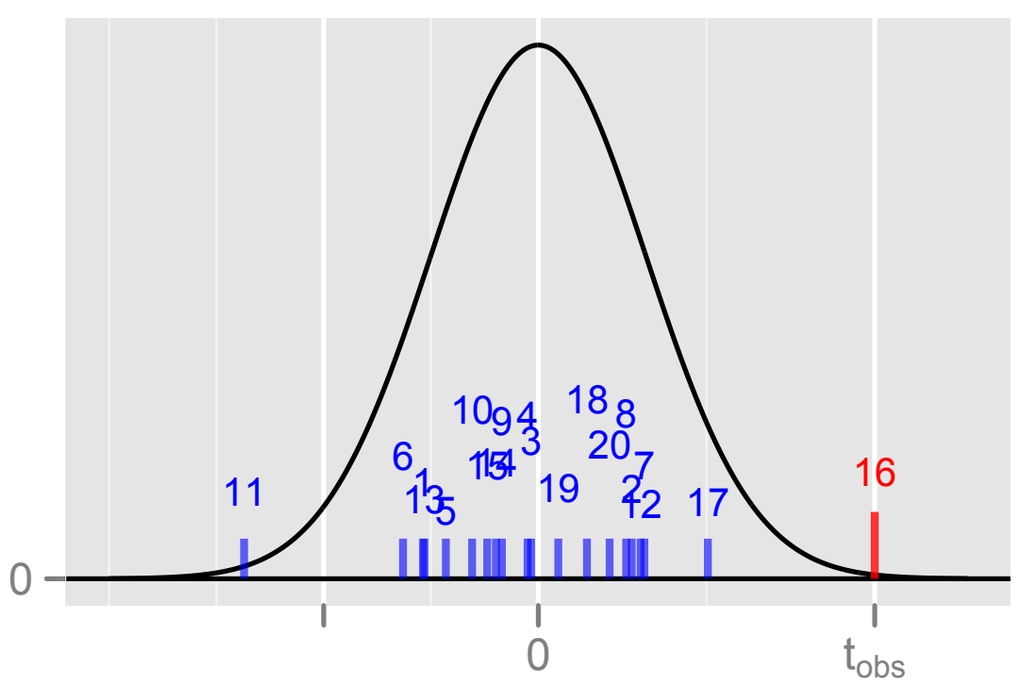
Source: Roy Chowdhury (2014)

Consideration ONE

Sampling distribution comparison is against a finite



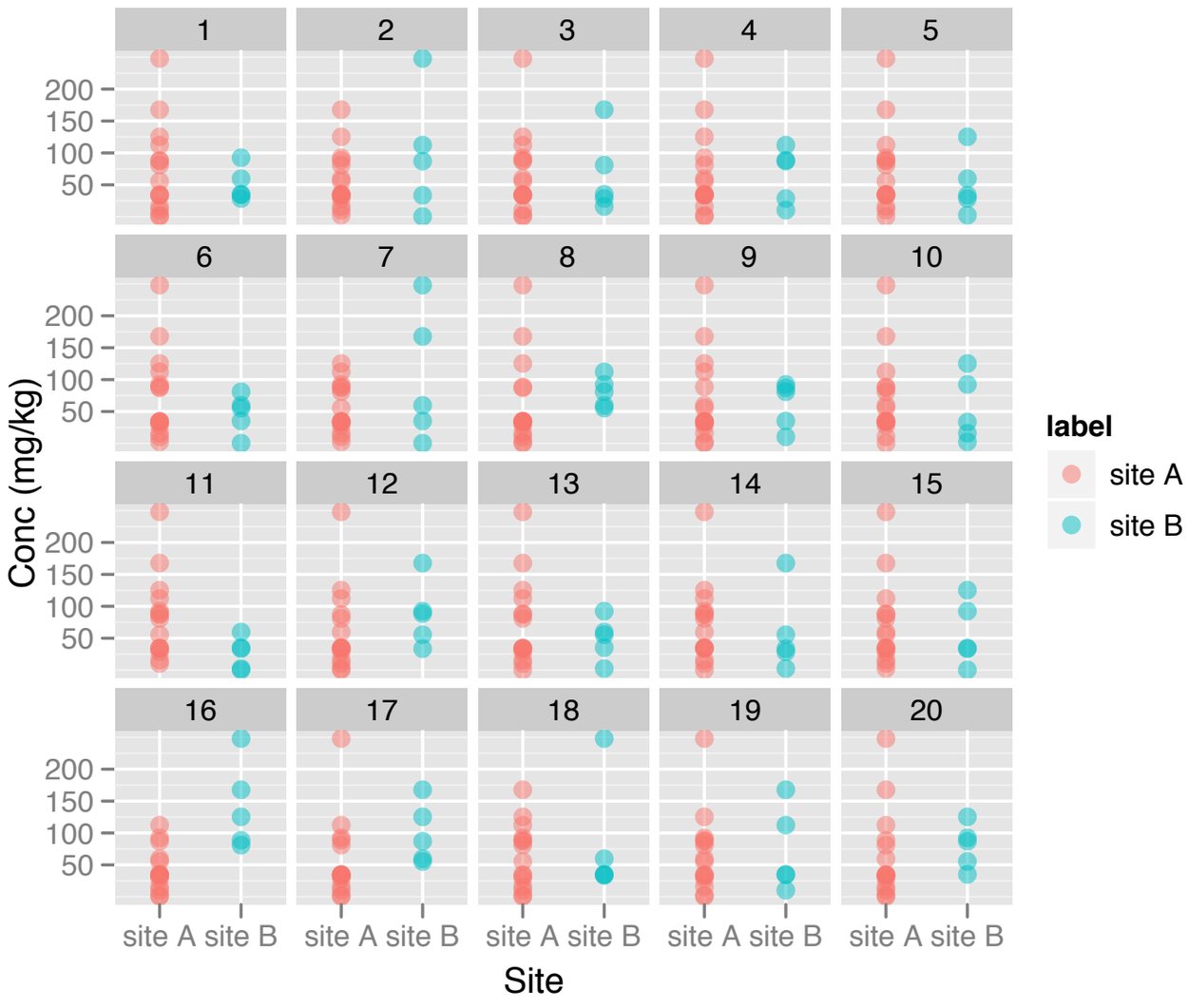


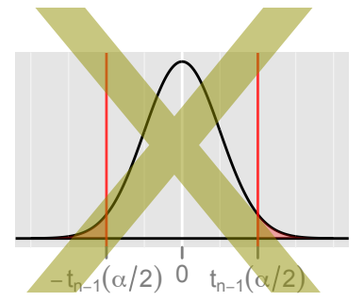


Source: Roy Chowdhury (2014)

Consideration ONE

Sampling distribution comparison is against a finite

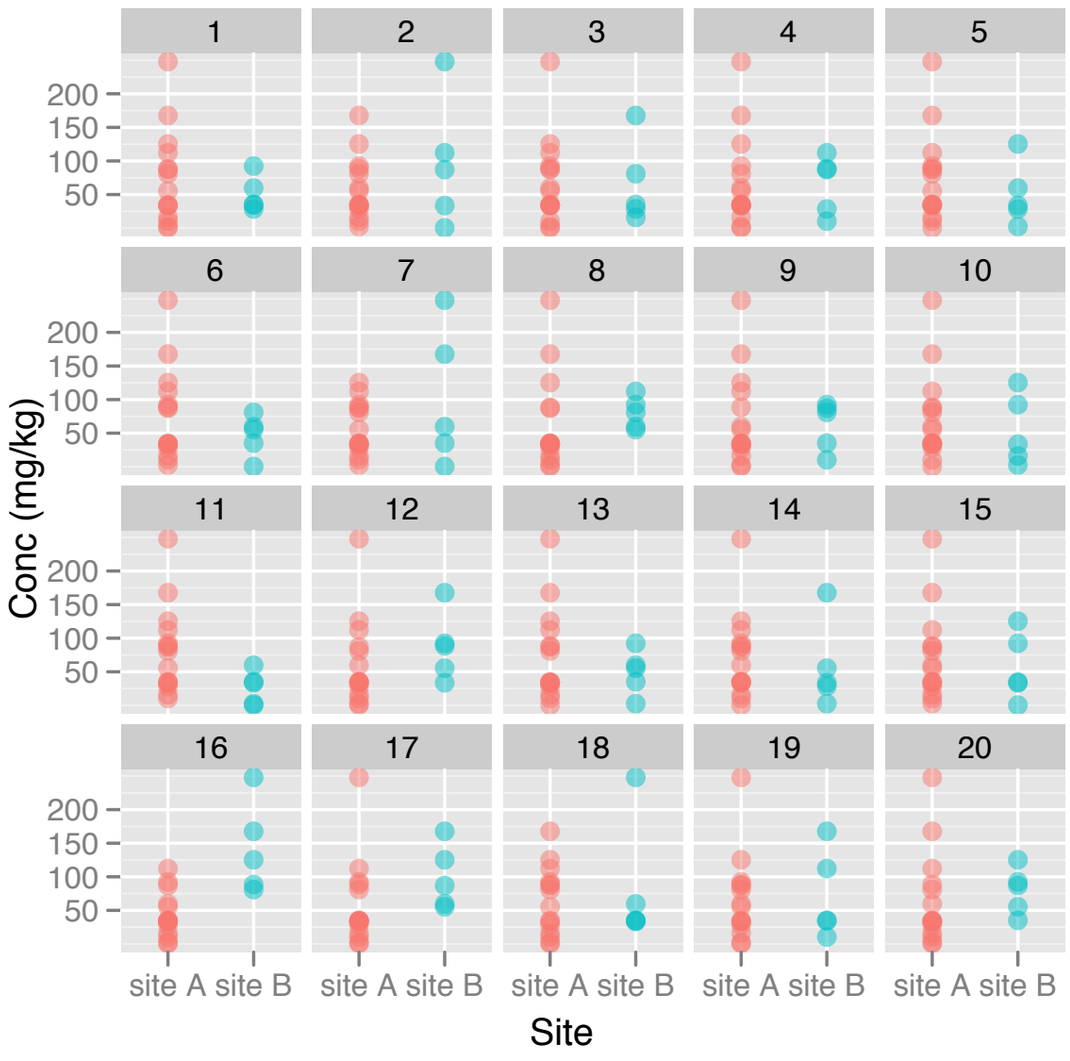


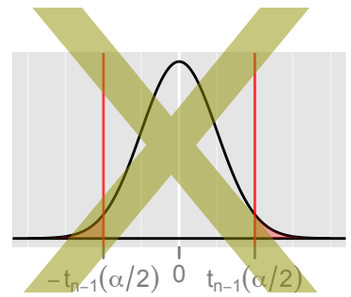


Source: Roy Chowdhury (2014)

Consideration ONE

Sampling distribution comparison is against a finite



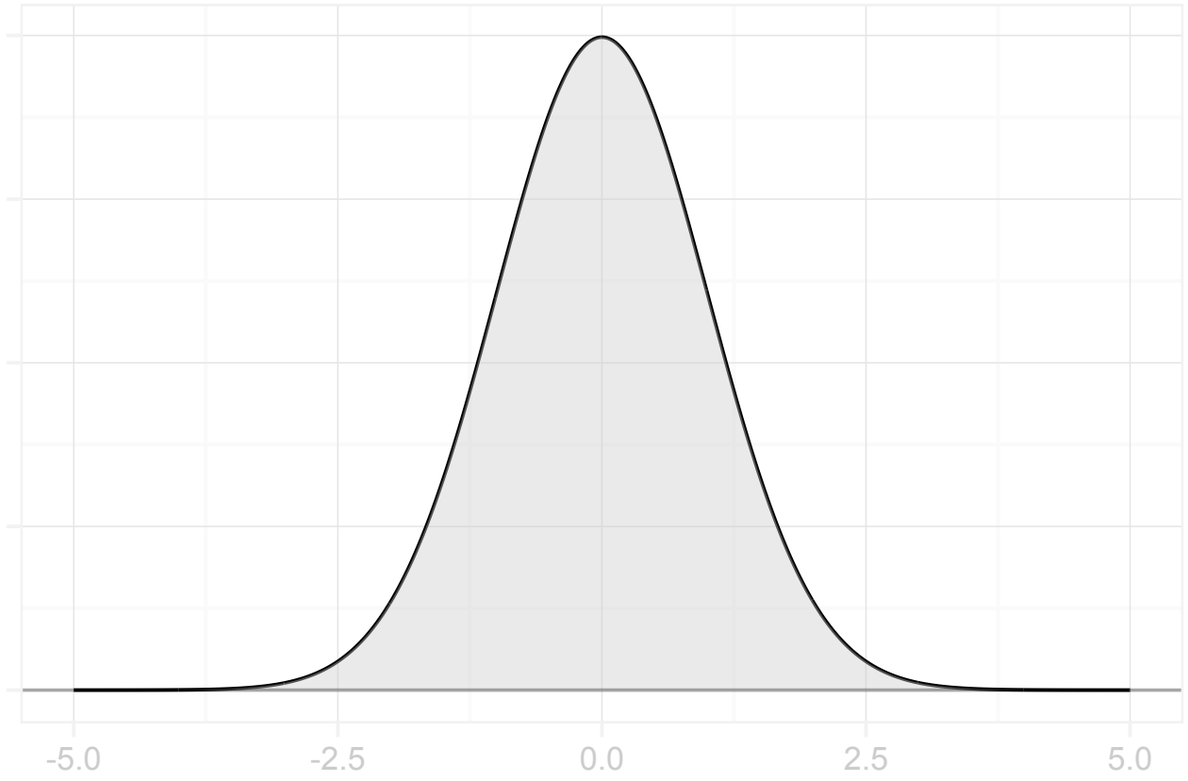


Measures of the QUALITY of a lineup

Source: Roy Chowdhury (2014)

Consideration ONE

KEEP IN MIND: In practice, graphics is being used when there is no quantification of a sampling distribution. All we have is (m-1) representatives from whatever

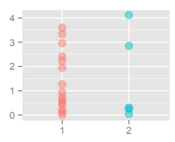


Consideration ONE

KEEP IN MIND: In practice, graphics is being used when there is no quantification of a sampling distribution. All we have is (m-1) representatives from whatever that distribution is.

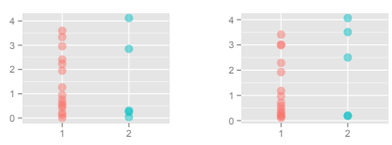
Consideration ONE

KEEP IN MIND: In practice, graphics is being used when there is no quantification of a sampling distribution. All we have is (m-1) representatives from whatever that distribution is.



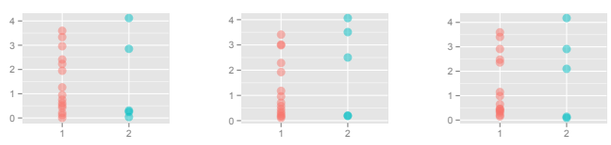
Consideration ONE

KEEP IN MIND: In practice, graphics is being used when there is no quantification of a sampling distribution. All we have is (m-1) representatives from whatever that distribution is.



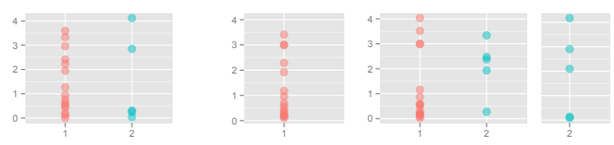
Consideration ONE

KEEP IN MIND: In practice, graphics is being used when there is no quantification of a sampling distribution. All we have is (m-1) representatives from whatever that distribution is.



Consideration ONE

KEEP IN MIND: In practice, graphics is being used when there is no quantification of a sampling distribution. All we have is (m-1) representatives from whatever that distribution is.



Consideration TWO

- What is the *p*-value?
- For one observer, the probability of randomly selecting the data plot is $^1/_m$, where m is the number of plots in the lineup.
- With multiple observers, the *p*-value is estimated by

$$P(X \geq x) = 1 - \textbf{\textit{Binom}}_{K,1/m}(x-1) = \sum_{i=x}^{K} \binom{K}{i} \left(\frac{1}{m}\right)^i \left(\frac{m-1}{m}\right)^{K-i}$$

Source: Majumder et al (2013) To appear

Consideration TWO

- What is the *p*-value?
- For one observer, the probability of randomly selecting the data plot is $^1/_m$, where m is the number of plots in the lineup.
- With multiple observers, the *p*-value is estimated by

Number of independent observers

$$P(X \geq x) = 1 - \textbf{\textit{Binom}}_{K,1/m}(x-1) = \sum_{i=x}^{K} \binom{K}{i} \left(\frac{1}{m}\right)^i \left(\frac{m-1}{m}\right)^{K-i}$$

Source: Majumder et al (2013) To appear

Consideration TWO

- What is the *p*-value?
- For one observer, the probability of randomly selecting the data plot is $^1/_m$, where m is the number of plots in the lineup.
- With multiple observers, the *p*-value is estimated by

$$P(X \geq x) = 1 - \textbf{\textit{Binom}}_{K,1/m}(x-1) = \sum_{i=x}^{K} \binom{K}{i} \left(\frac{1}{m}\right)^i \left(\frac{m-1}{m}\right)^{K-i}$$

Source: Majumder et al (2013) To appear

Consideration TWO

- What is the *p*-value?
- For one observer, the probability of randomly selecting the data plot is $^1/_m$, where m is the number of plots in the lineup.
- With multiple observers, the *p*-value is estimated by

$$P(X \geq x) = 1 - Binom_{K,1/m}(x-1) = \sum_{i=x}^{K} \binom{K}{i} \left(\frac{1}{m}\right)^i \left(\frac{m-1}{m}\right)^{K-i}$$

Number of observers choosing data plot

Source: Majumder et al (2013) To appear

Consideration TWO

- What is the *p*-value?
- For one observer, the probability of randomly selecting the data plot is $^1/_m$, where m is the number of plots in the lineup.
- With multiple observers, the *p*-value is estimated by

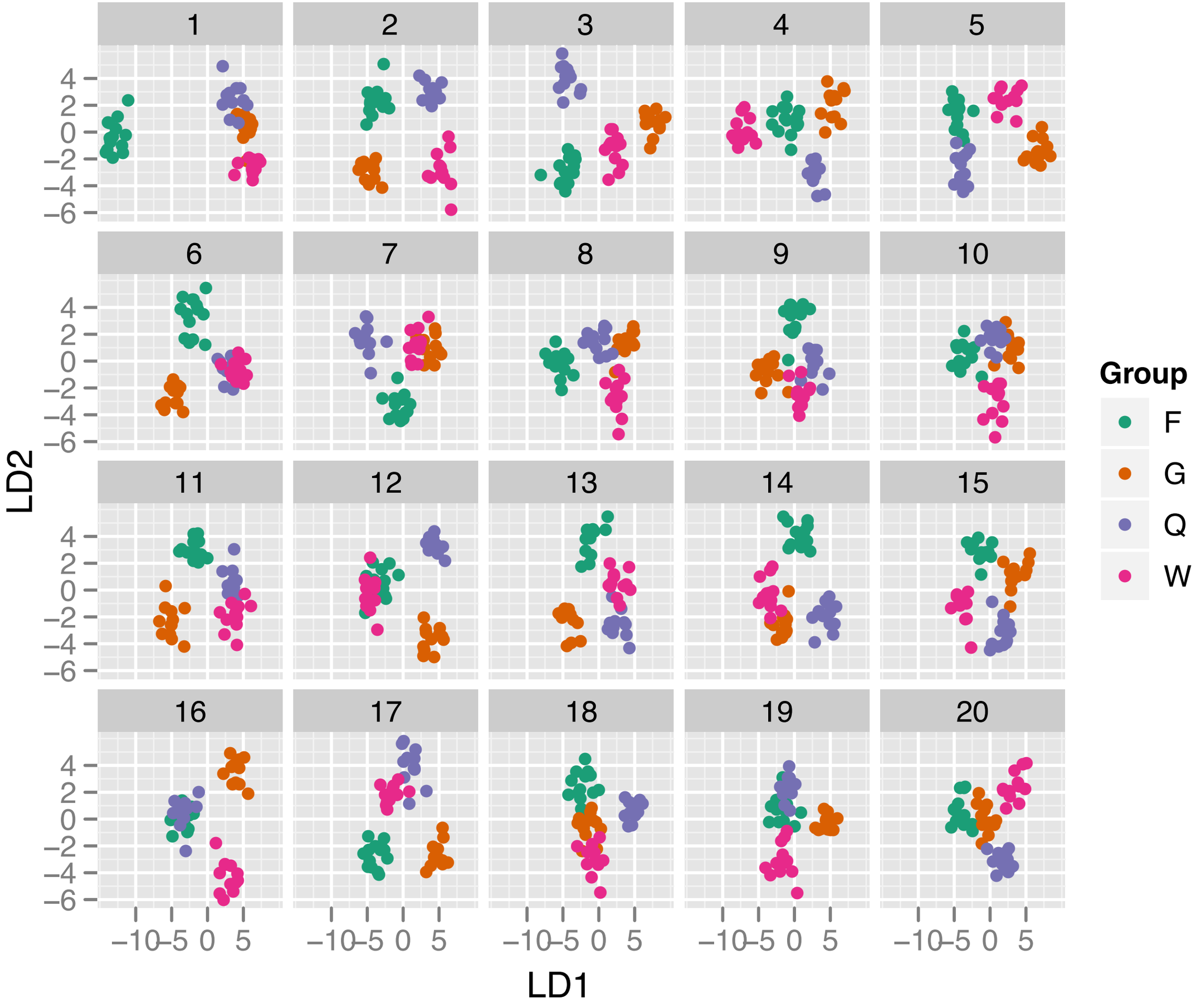
$$P(X \geq x) = 1 - \textbf{\textit{Binom}}_{K,1/m}(x-1) = \sum_{i=x}^{K} \binom{K}{i} \left(\frac{1}{m}\right)^i \left(\frac{m-1}{m}\right)^{K-i}$$

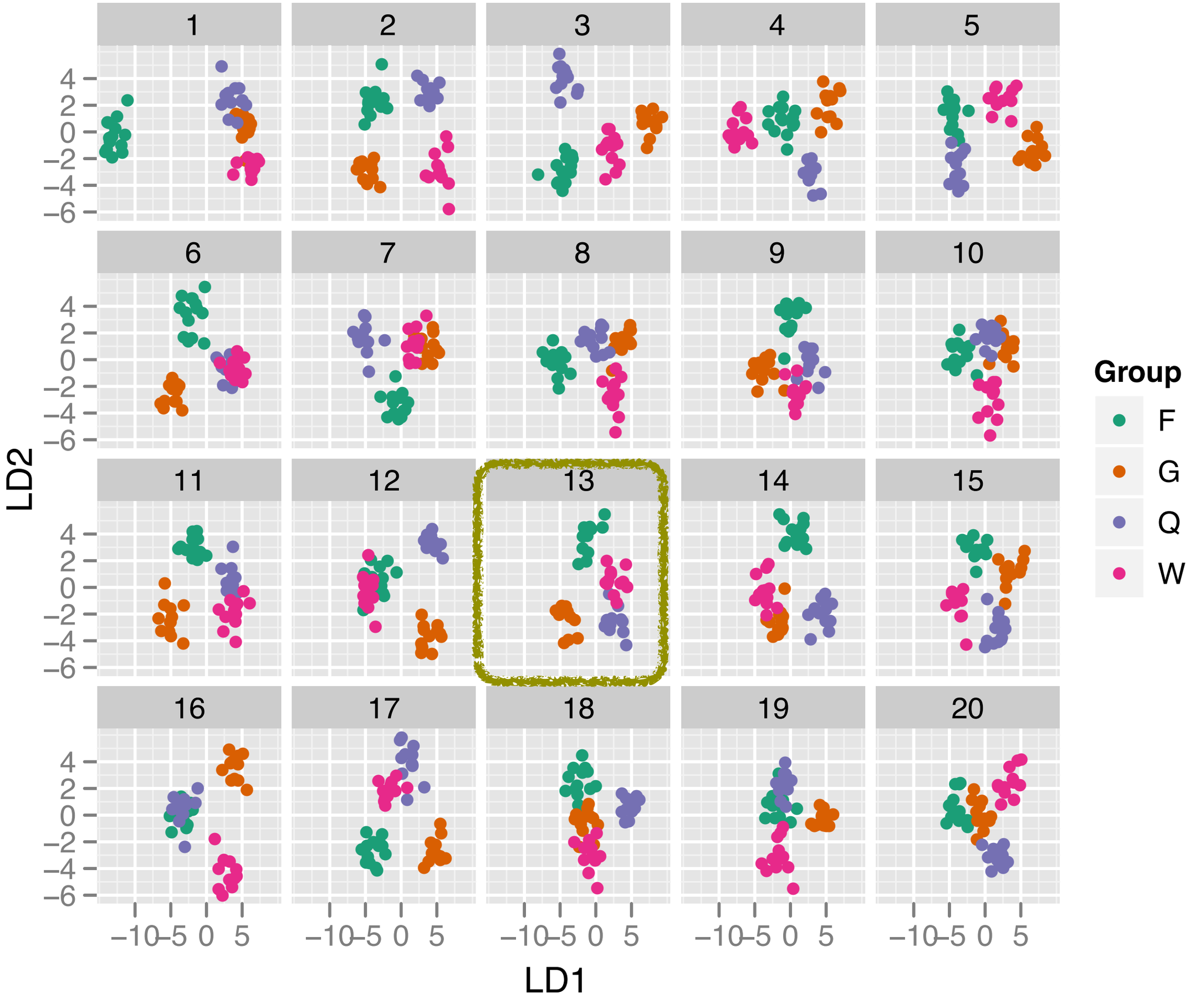
Source: Majumder et al (2013) To appear

Consideration THREE

- What is the power of the test?
- There is a choice of type of plot to use. Some will be more optimal than others.
- Signal strength will be defined as “proportion of observers who identify the data plot”.
- Enables the comparison of different plot designs.
- Signal strength equals power, when only plot design changes.

Source: Majumder et al (2013) To appear
Hofmann et al (2012) InfoVis ’12 Proc


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
4
2
0
−2
−4
−6
−10
−5
0
5
LD2
LD1
Group
F
G
Q
W


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
LD2
LD1
4
2
0
−2
−4
−6
−10
−5
0
5
Group
F
G
Q
W

What we learn

- Wasps data is no different from random assignment of species label
- Difference between groups was due to sparseness of high-dimensional space

Your Turn

- Jot down the number of the plot

1	2	3	4	5
6	7	8	9	10
11	12	13	14	15
16	17	18	19	20

- Write down a reason for your choice
- And rate how confident that you picked the data on a scale of 1 (very sure) to 5 (don't have a clue)

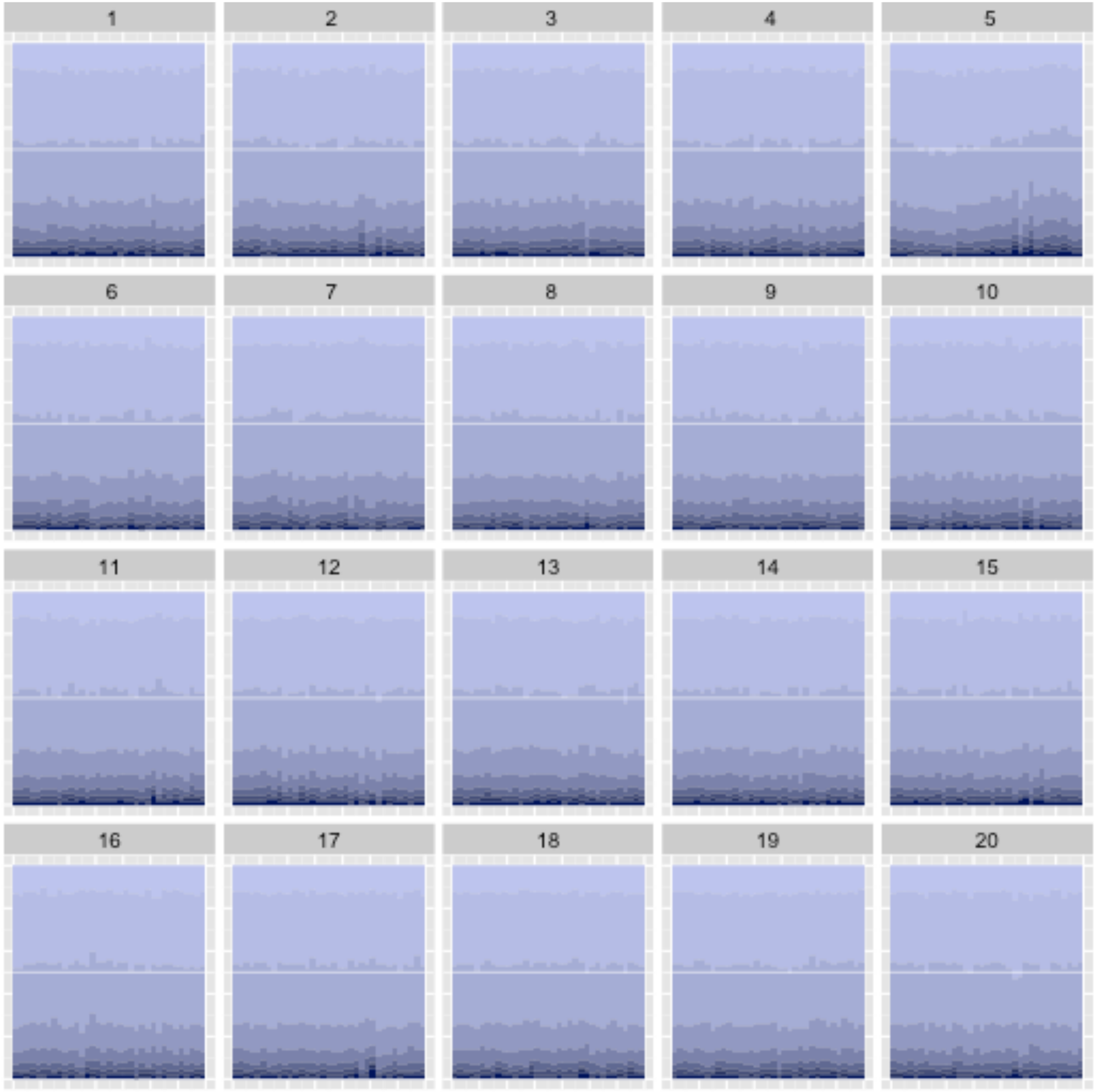
Your Turn

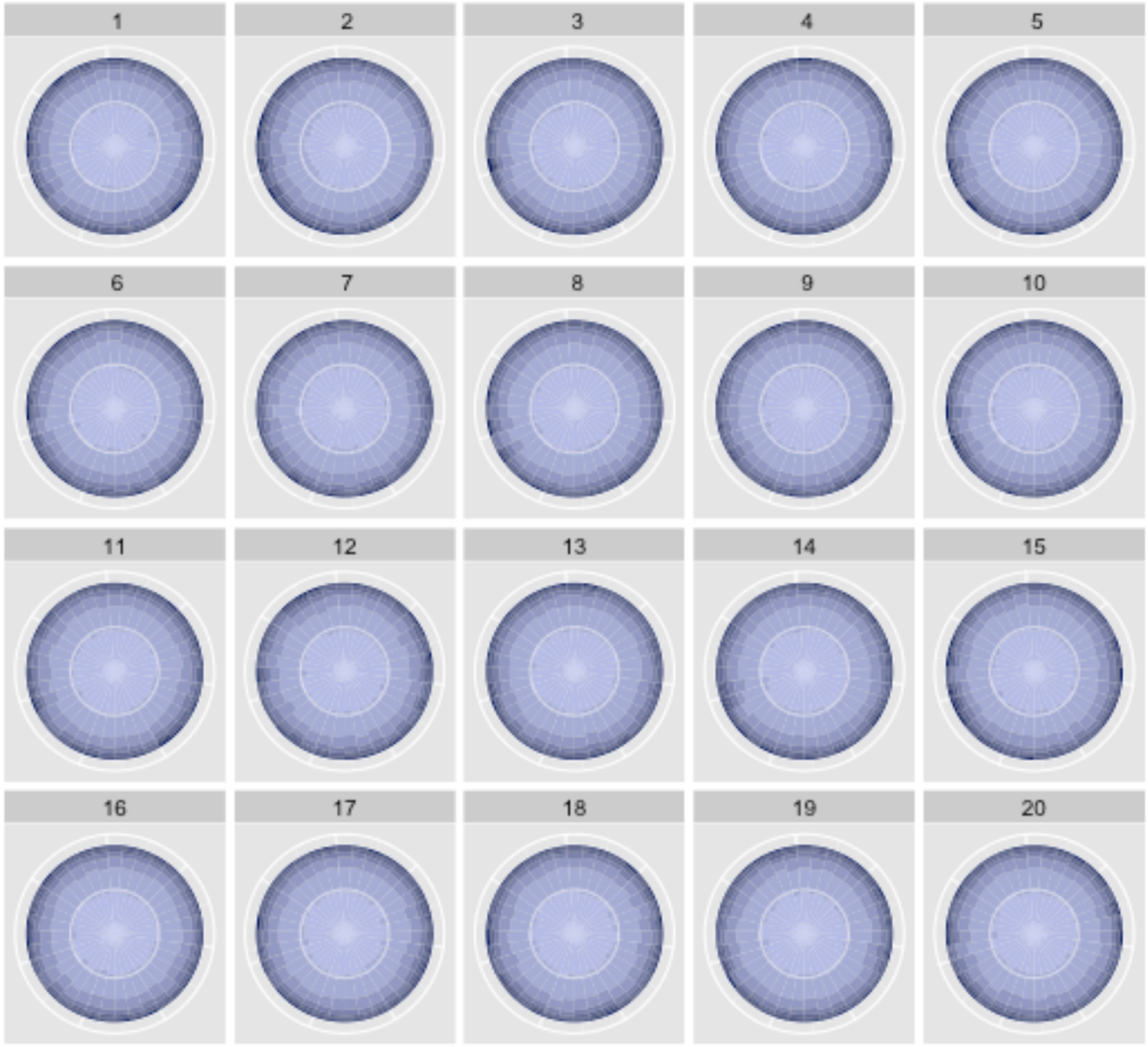
For the following lineups of 20 plots, pick the ONE plot that is most different from the others

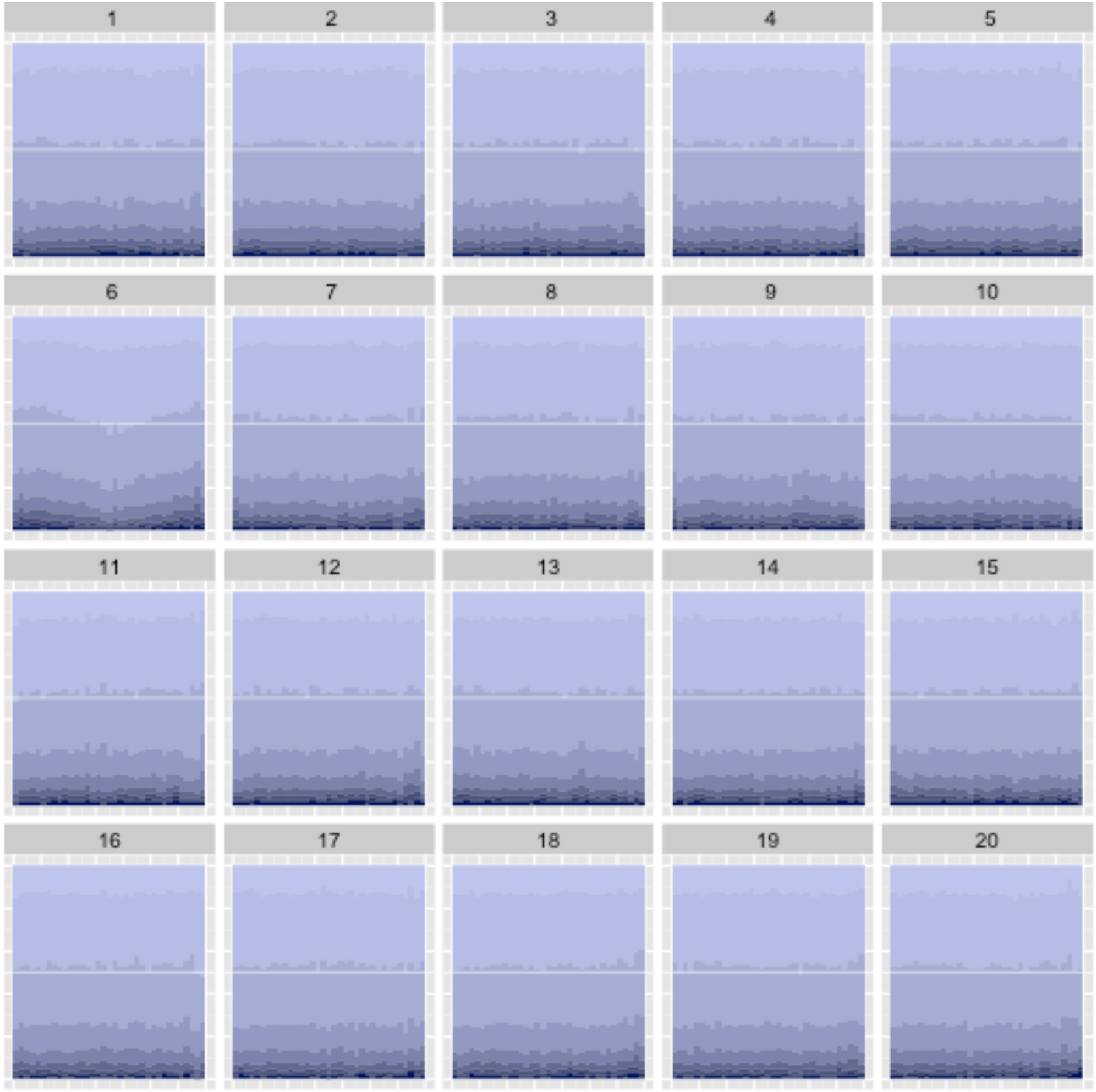
- Jot down the number of the plot

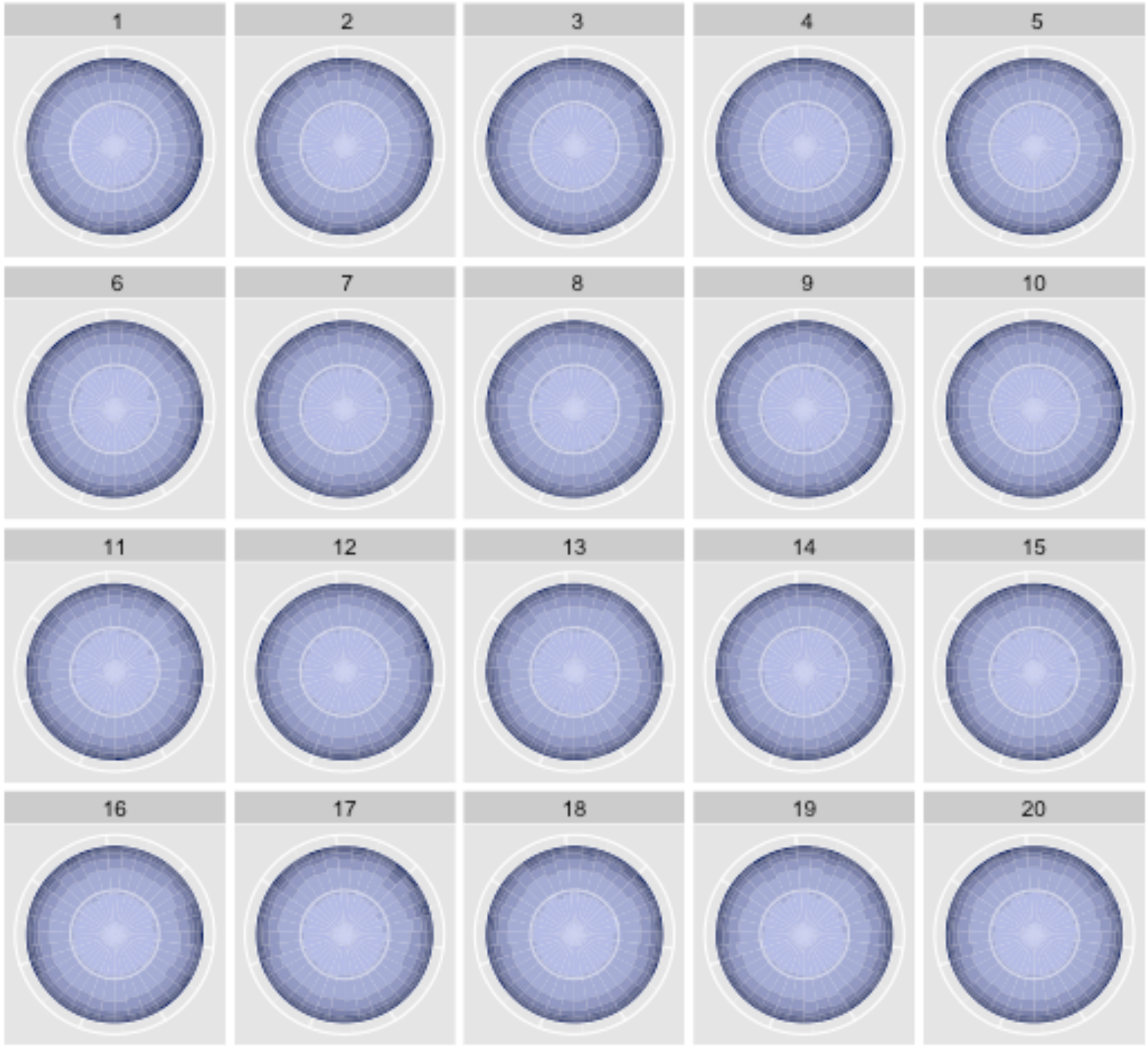
1	2	3	4	5
6	7	8	9	10
11	12	13	14	15
16	17	18	19	20

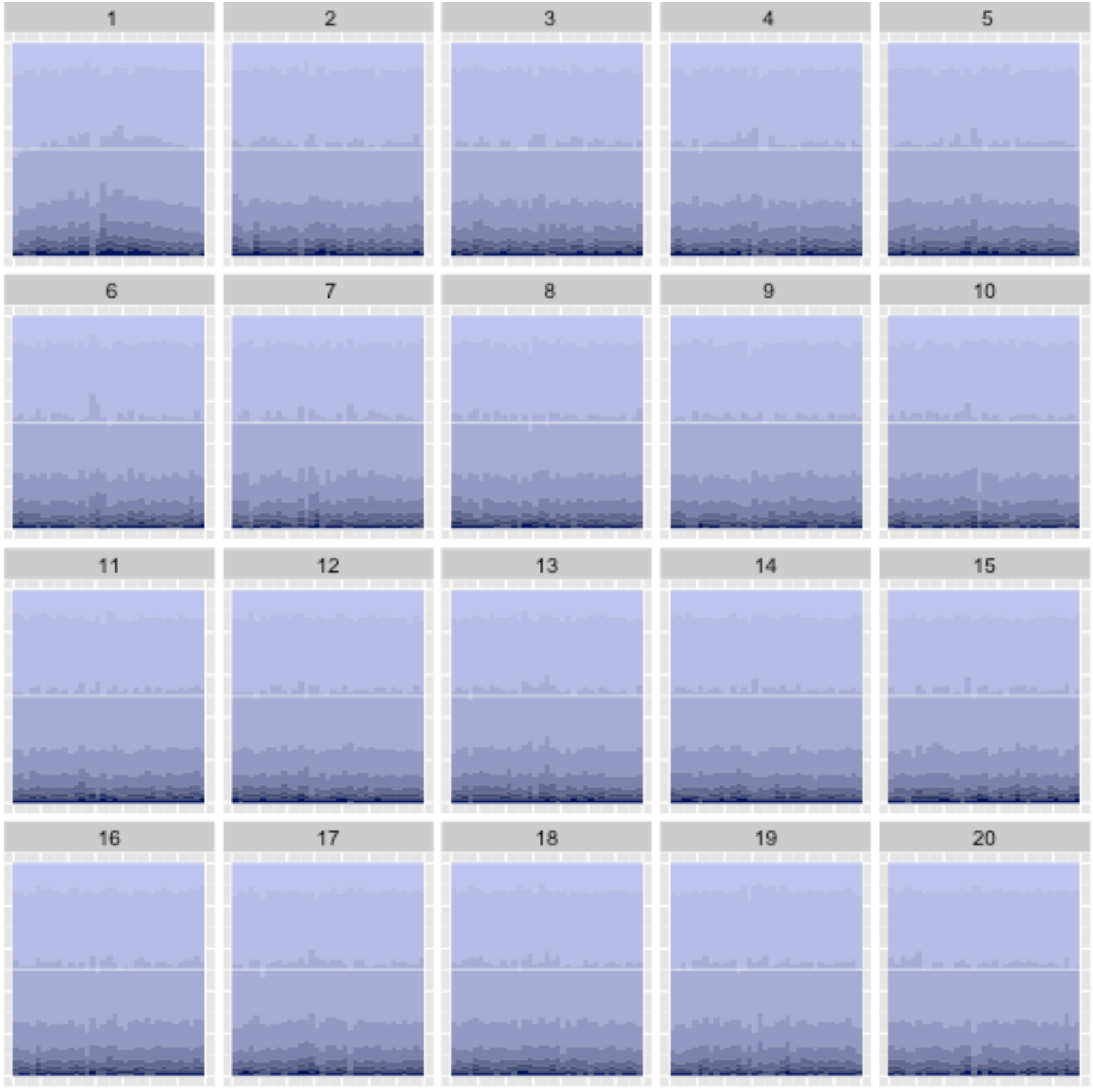
- Write down a reason for your choice
- And rate how confident that you picked the data on a scale of 1 (very sure) to 5 (don't have a clue)

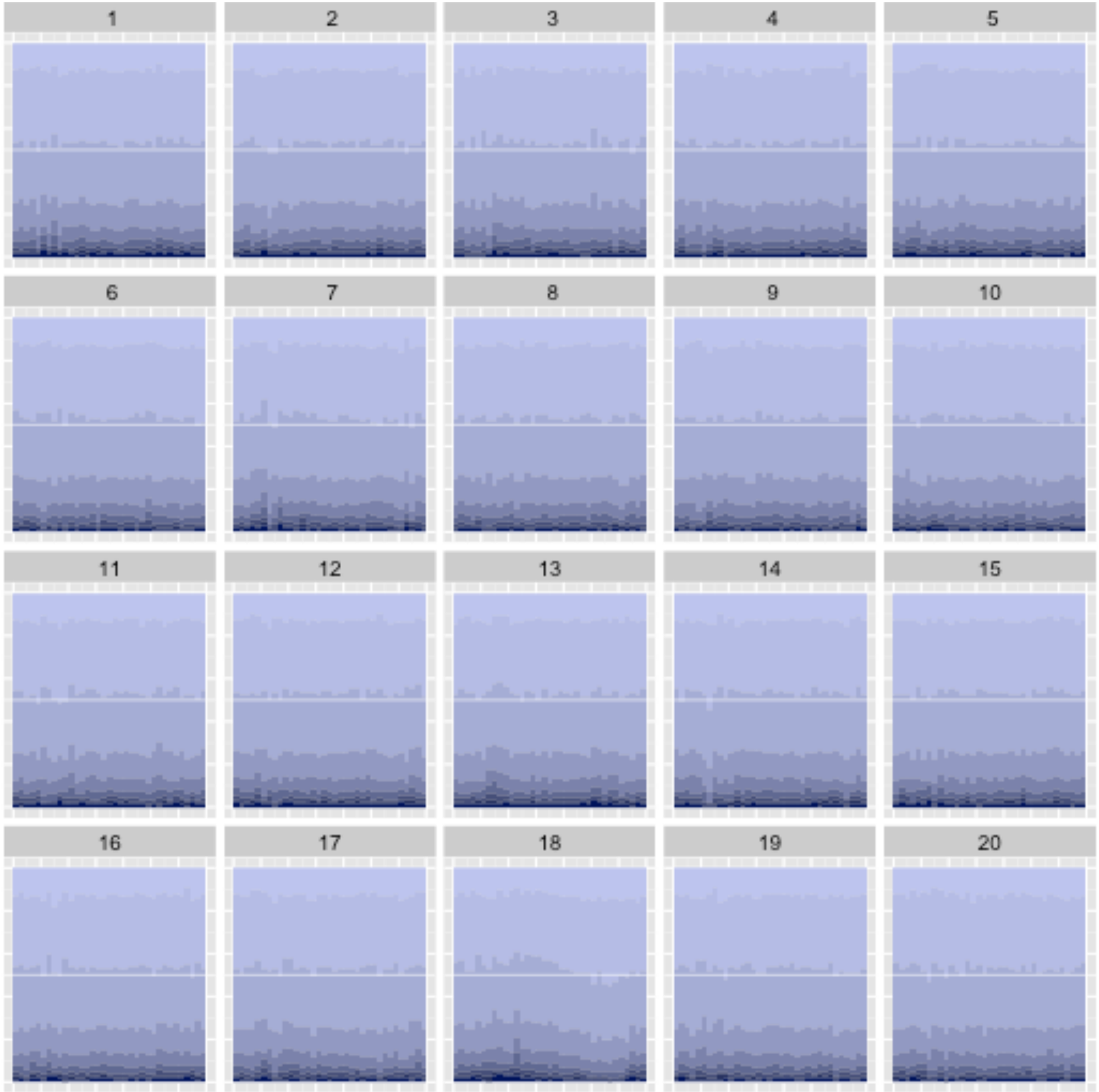

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20

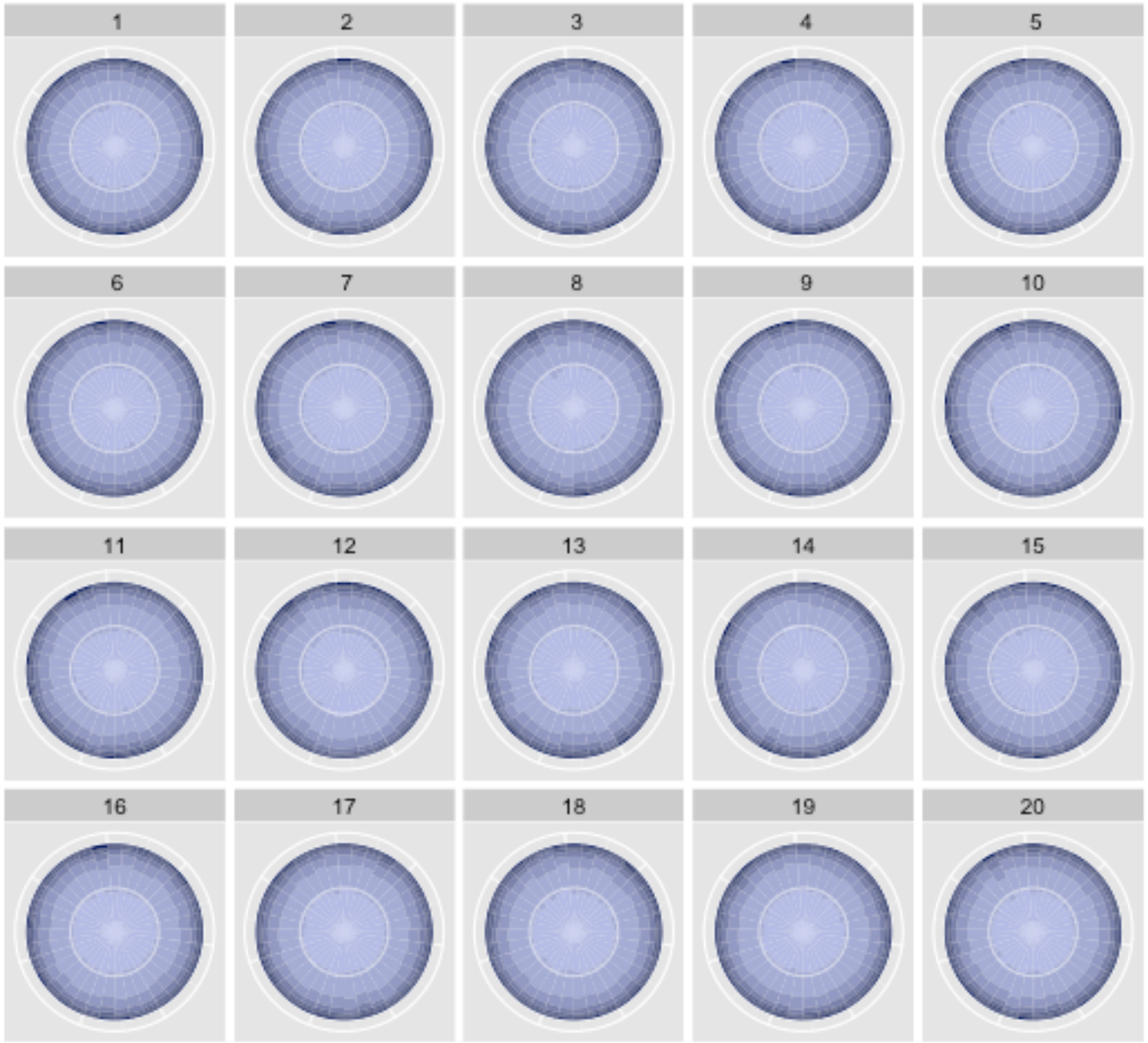

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20

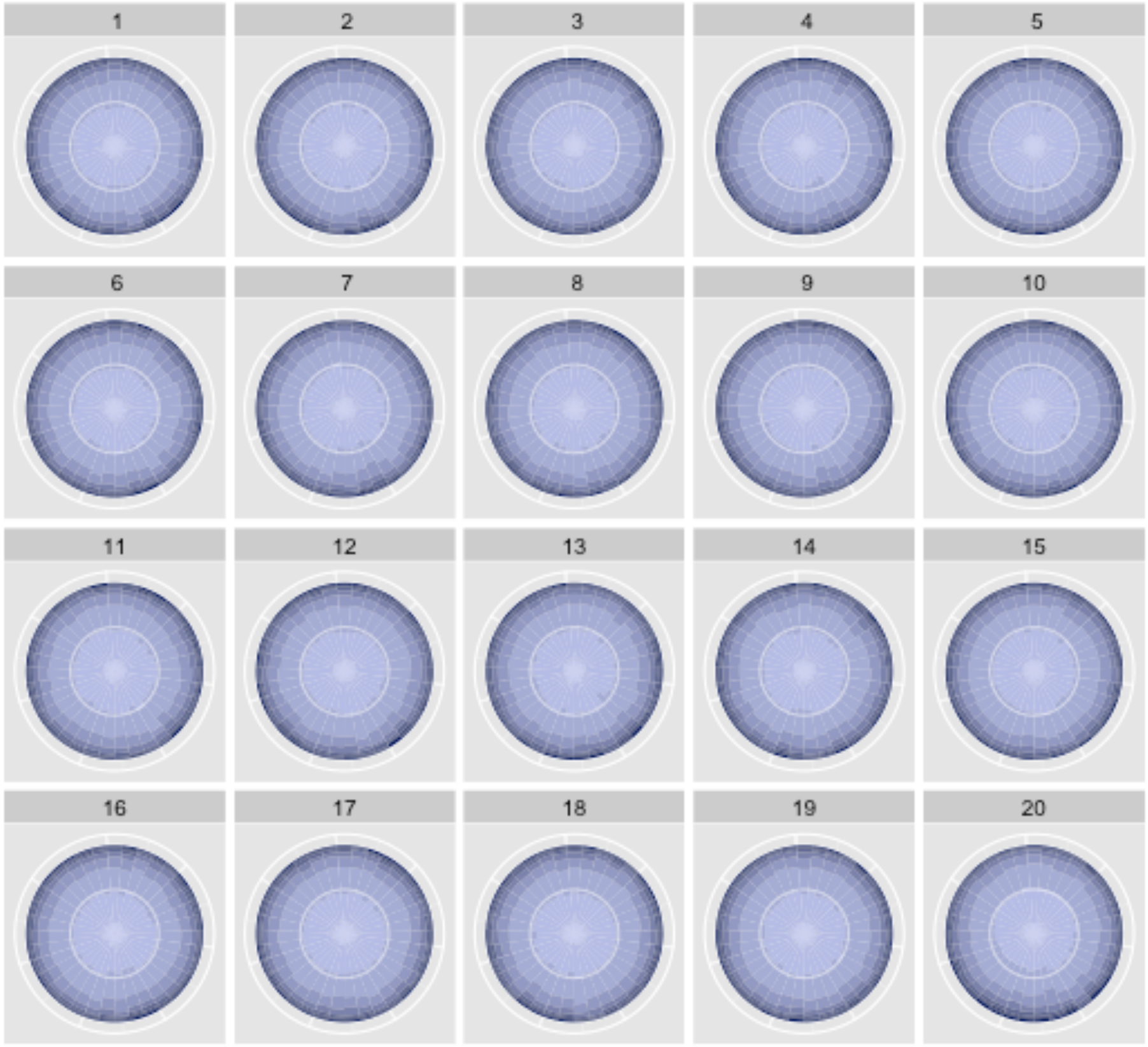

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20

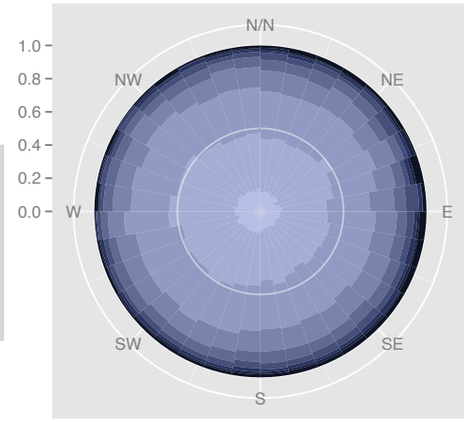

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20

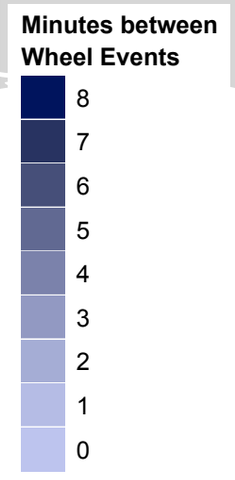

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20

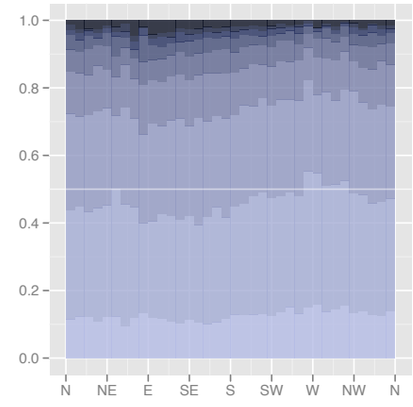

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20

	Lineup	# Correct	Reason	Confidence
5	1			
2	2			
6	3			
1	4			
1	5			
18	6			
12	7			
20	8			

Study







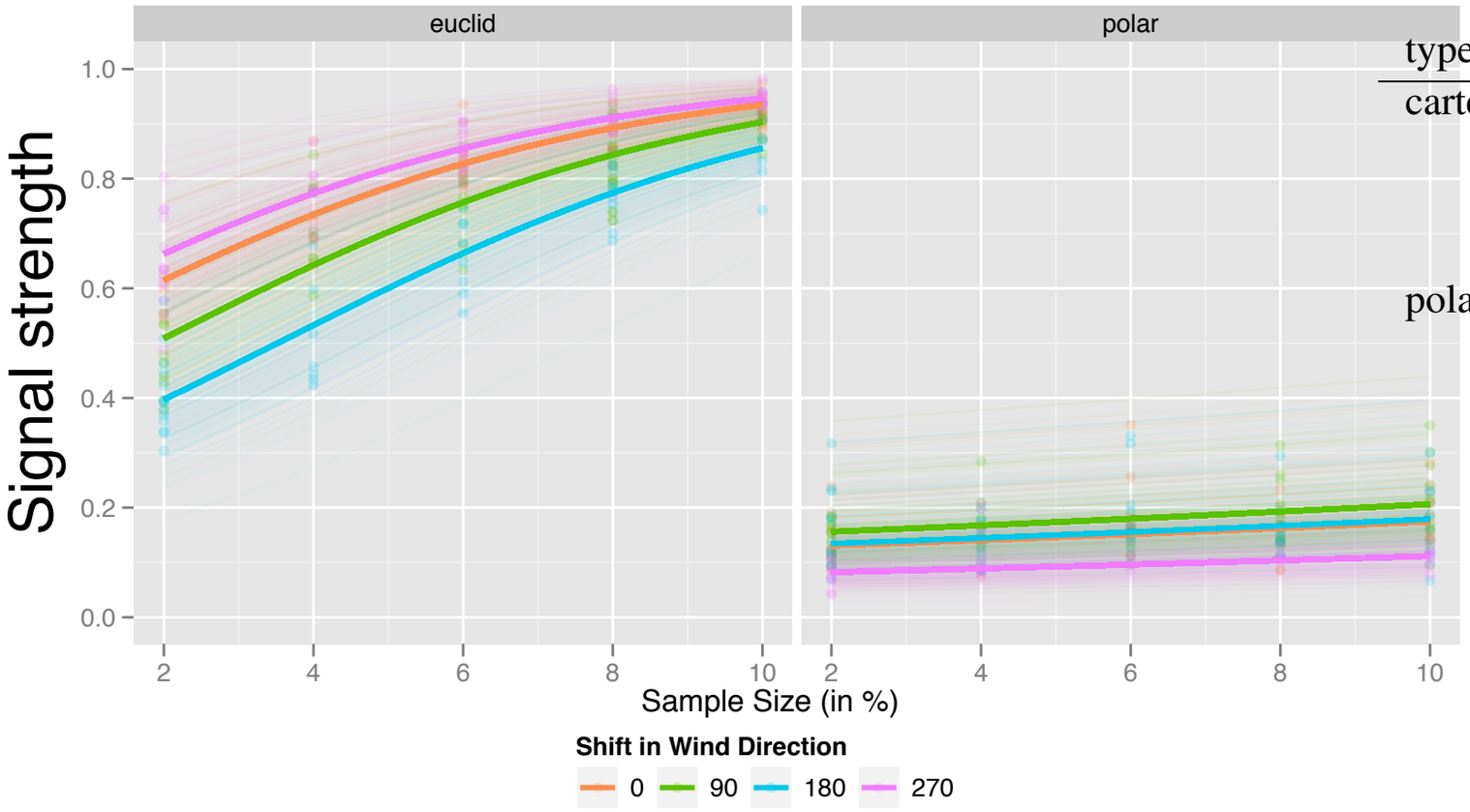
- Examine wind direction and airport efficiency.

H_0 : *wind direction has no effect on efficiency*

against the alternative H_a : *wind direction does have an effect.*

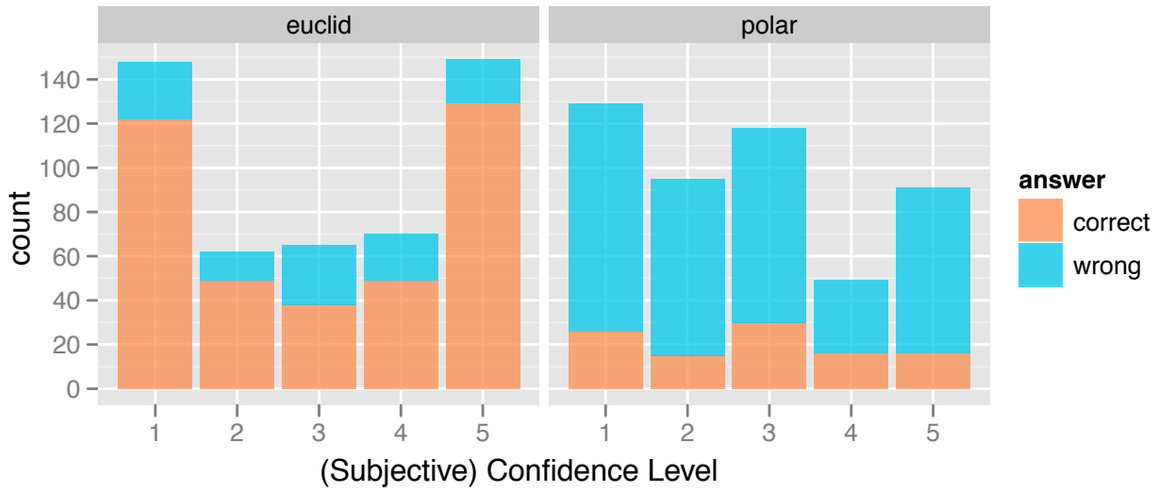
- Decide on best display: (conditional) wind rose charts or bar charts, where each of 36 wind directions the percentage of flights falling into one-minute intervals between successive flights, from zero minutes to eight minutes or more is shown in color scale.

Results of full study



type of chart		sample size 2	4	6	8	10	24
cartesian	with	12/**24**	18/**27**	42/**44**	33/**41**	39/**43**	46/**47**
		(0.50)	(0.67)	(0.95)	(0.80)	(0.91)	(0.98)
	without	13/**41**	24/**32**	39/**44**	47/**51**	40/**45**	34/**37**
		(0.32)	(0.75)	(0.89)	(0.92)	(0.89)	(0.92)
polar	with	13/**49**	9/**43**	9/**39**	7/**35**	8/**40**	13/**38**
		(0.27)	(0.21)	(0.23)	(0.20)	(0.20)	(0.34)
	without	6/**51**	4/**34**	5/**34**	6/**39**	11/**43**	12/**37**
		(0.12)	(0.12)	(0.15)	(0.15)	(0.26)	(0.32)

Reported confidence in their choice on plot

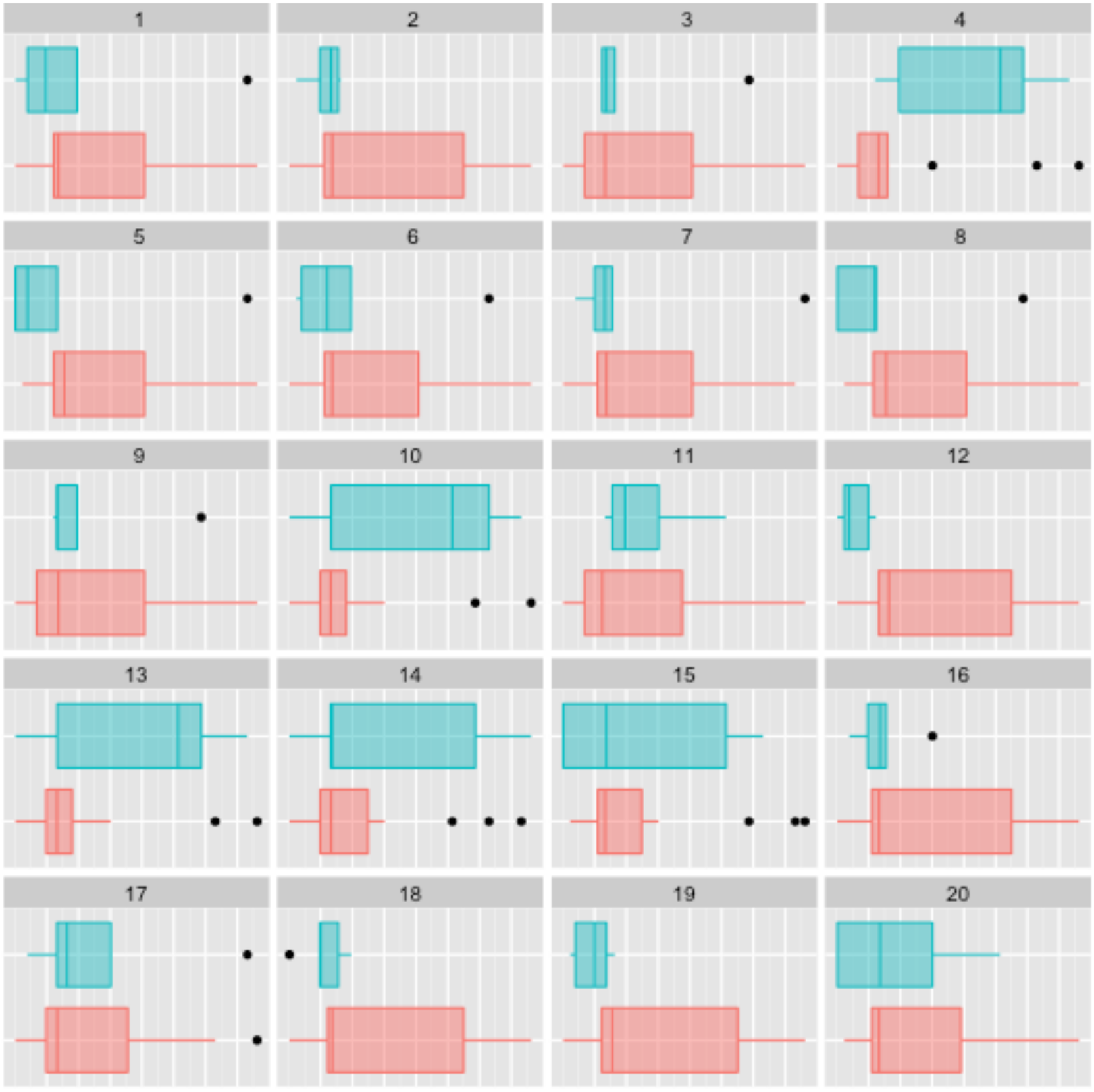


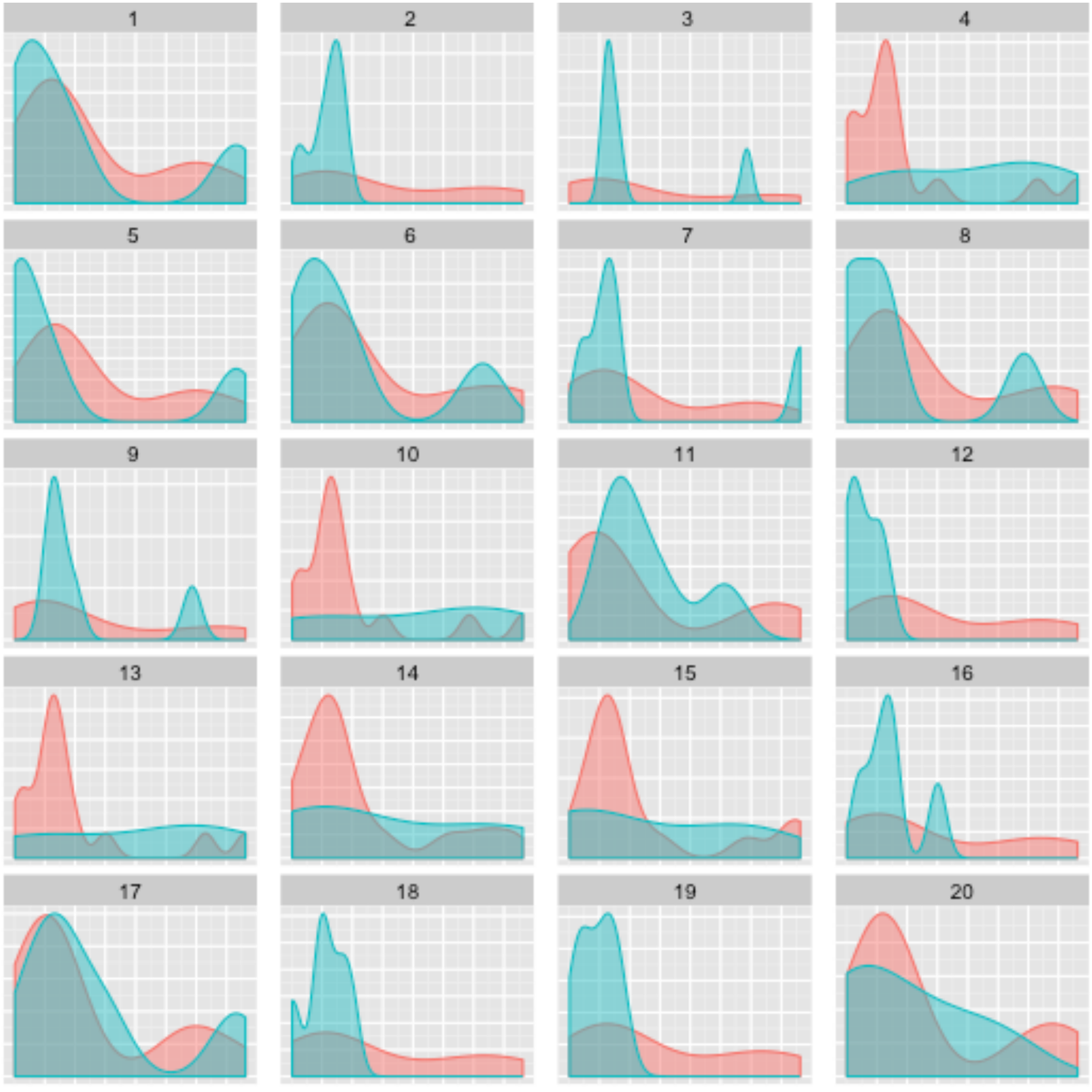
A similar study

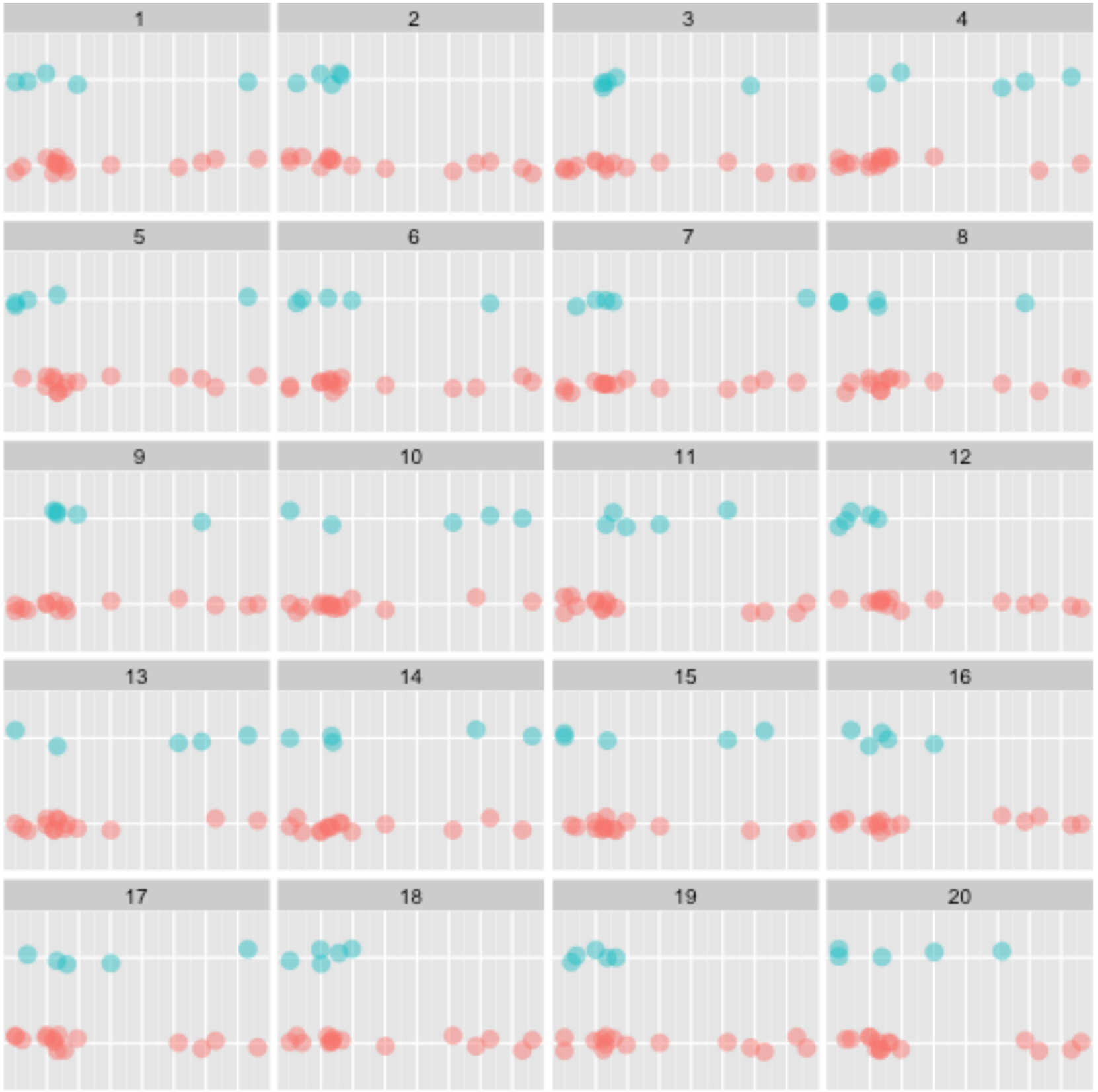
MOTIVATION: A very small data set of chemical concentrations taken from a superfund clean up site (5 values) , compared with samples taken from a normal site (15 values).

Can we see a difference between the two groups, using a side-by-side dotplot? Are side-by-side dotplots better for comparing groups, or side-by-side boxplots, or stacked histograms or density plots?

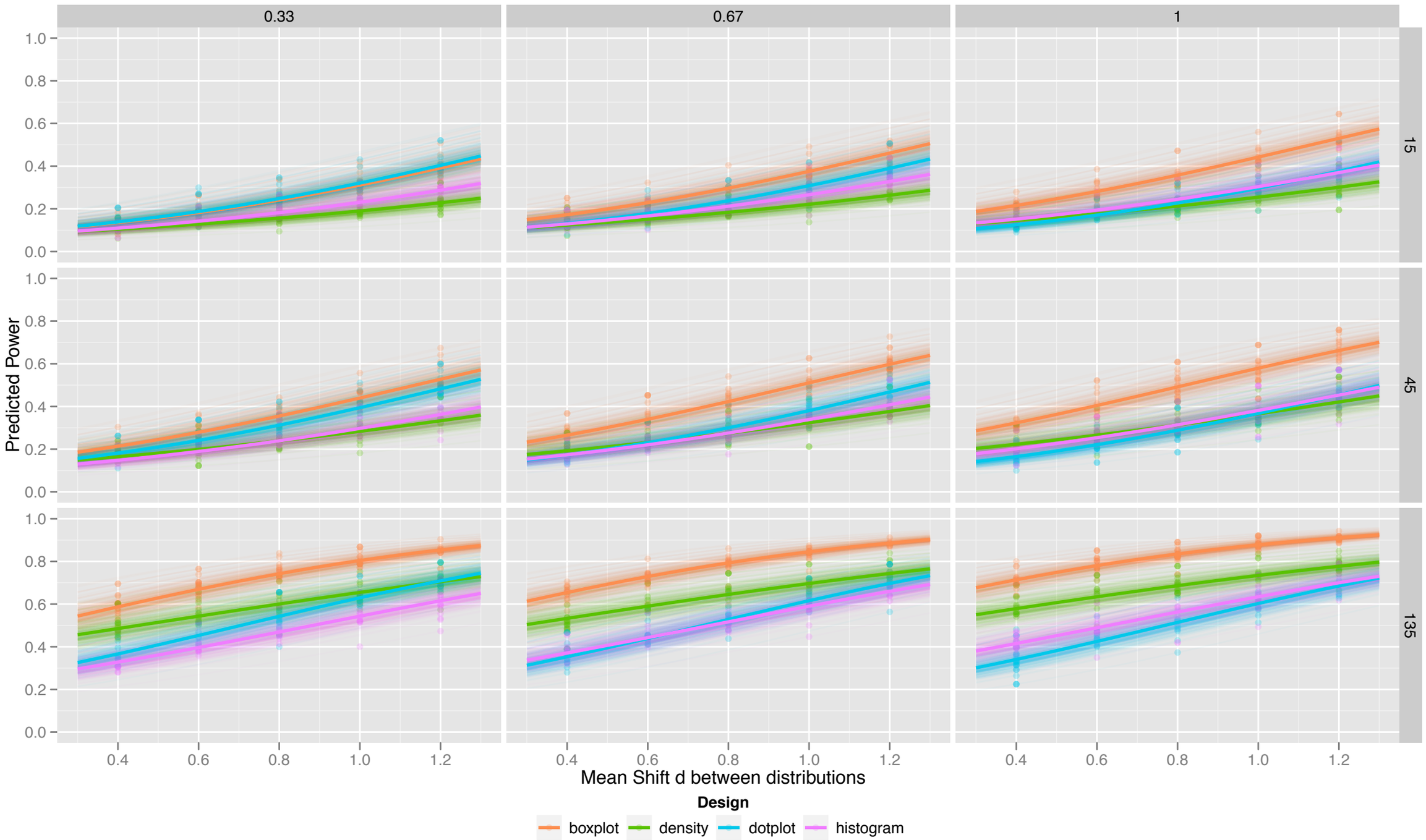
In which group is the blue group further to the right?


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20



Boxplots beats all other plot designs, except for really small data sets.

Process

1. Decide on appropriate plot of the data, using good graphical principles.
2. Make the lineup before you have seen the actual data plot.
3. Pick the plot that is different from the rest.
4. If you have already seen the plot of the data, you can show the lineup to someone who hasn't, and use their results.
5. Services like Amazon's Mechanical Turk allow employing independent observers, from a broad cross-section of society.
6. (We are not doing invalid post-hoc testing.)

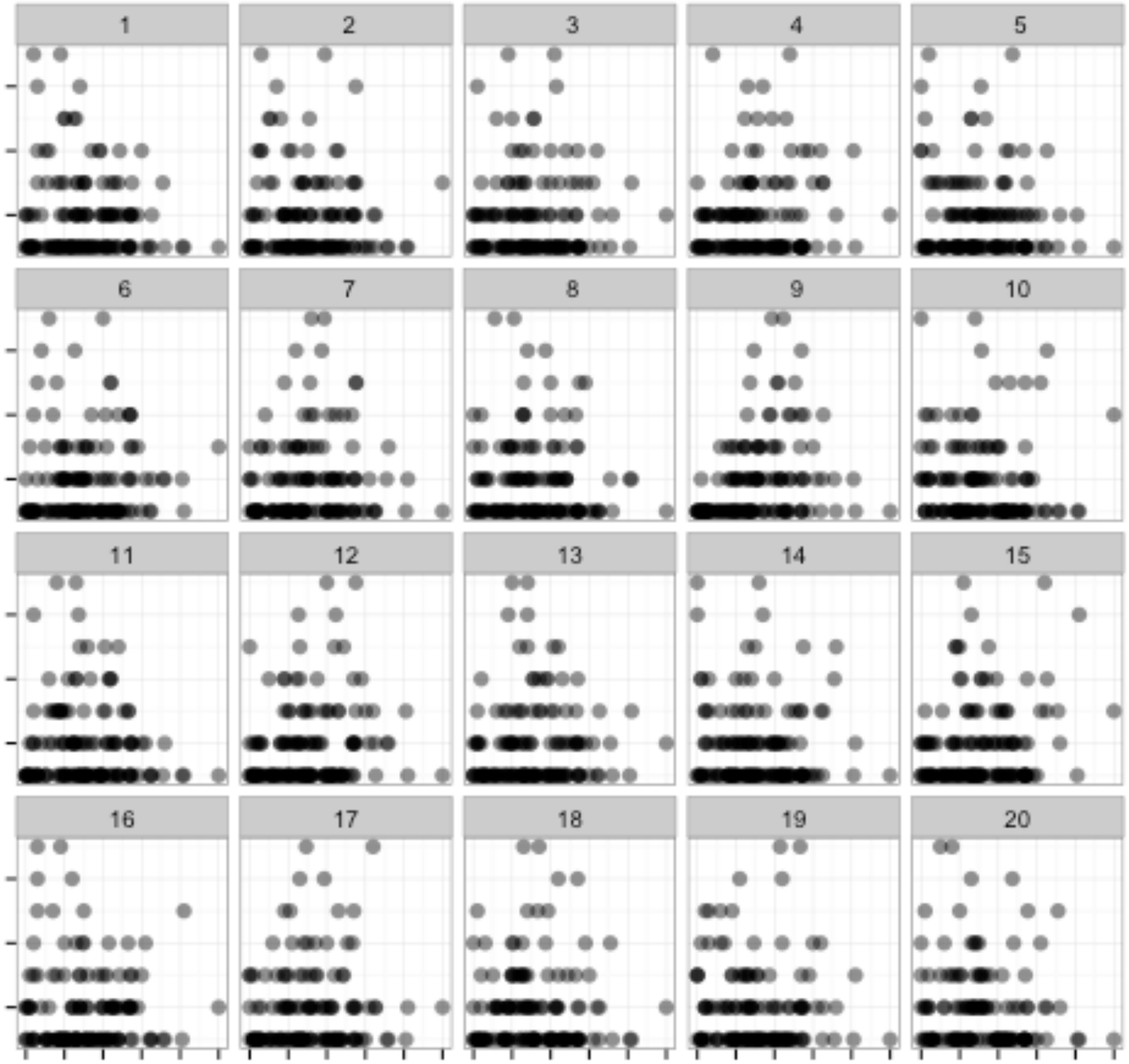
nullabor package

- Builds on the ggplot2 package for making data plots.
- Generate the lineups automatically, so that you see this before you see the plot of the data.
- Encrypts the location of the actual plot, for you to decrypt when you're ready.

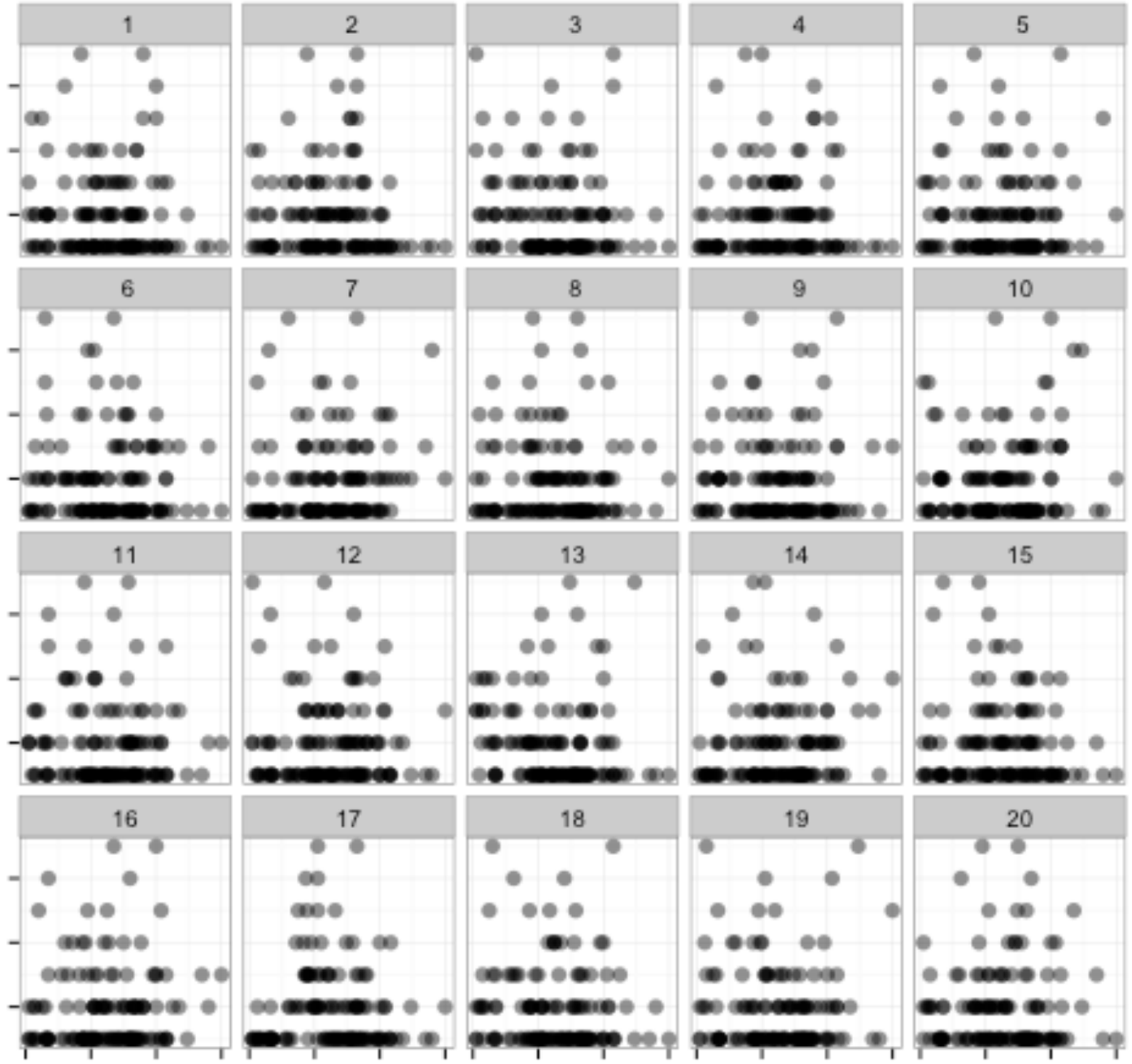
Tennis statistics

- The relationships between round and performance statistic are not regular.
- Simple linear model may not pick up if there is a relationship between the variables.
- Lineups can be used to determine if there is a ***real*** relationship.
- Permutation of "round" label is used to generate nulls

Ready?



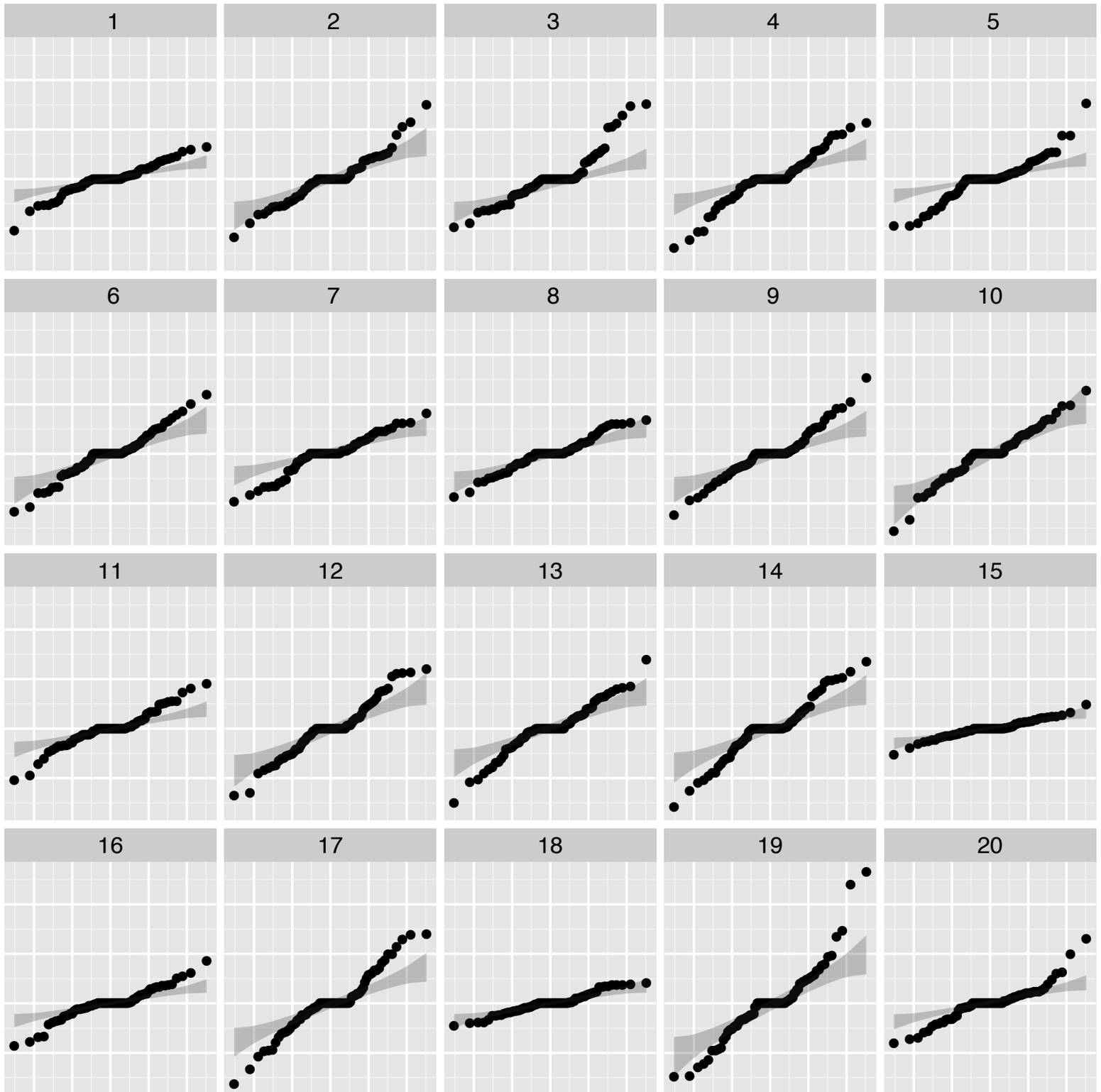
decrypt("Y25b yGKG Uu I1OUKU1u Xj")



decrypt("Y25b yGKG Uu I1OUKU1u Xj")

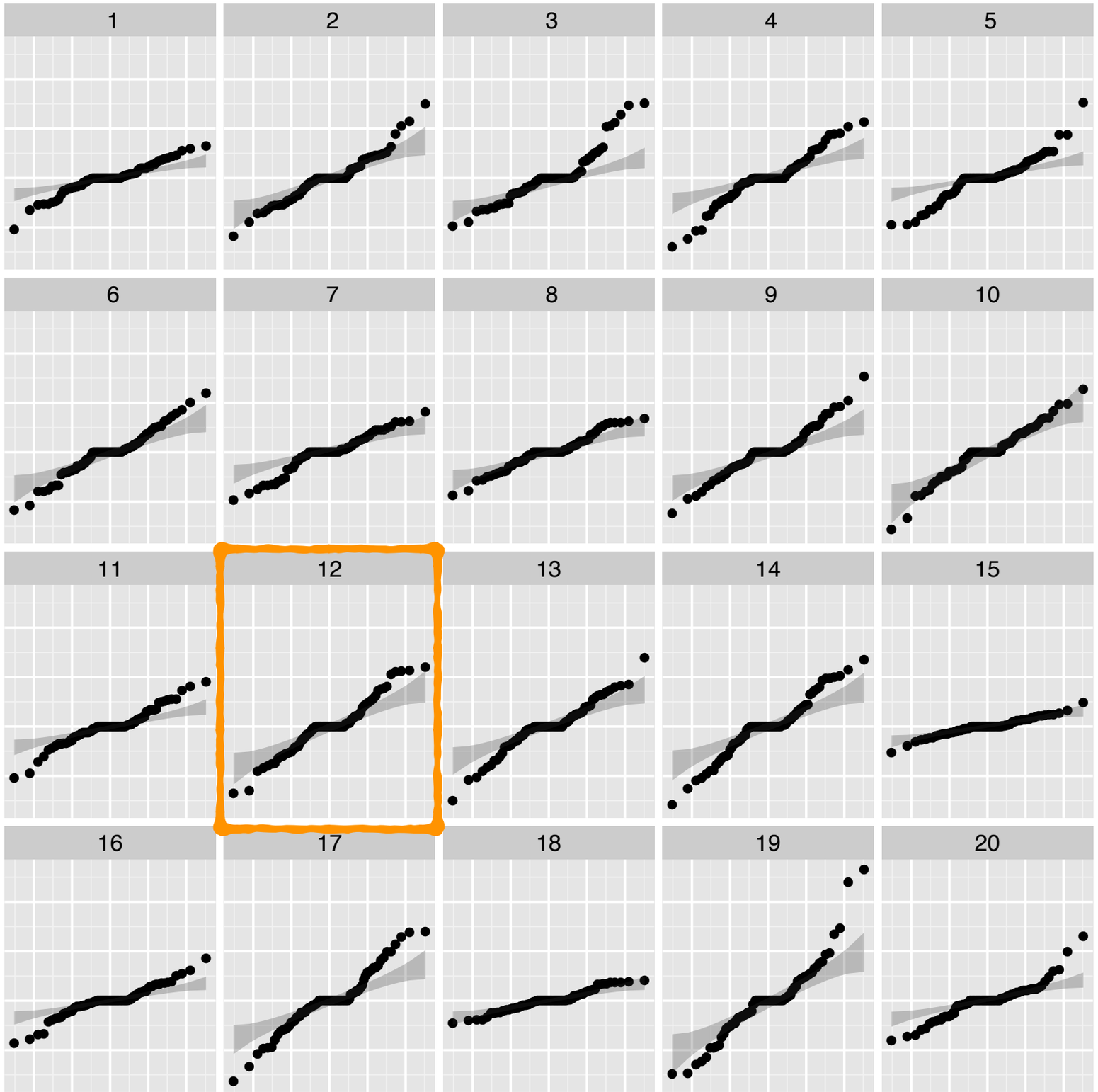
Use in our group

- Residual plots to diagnose hierarchical linear models
- Is there any structure in my RNA seq data?
- Does the false discover rate (FDR) give a good cutoff of the significant genes (RNA-seq)?



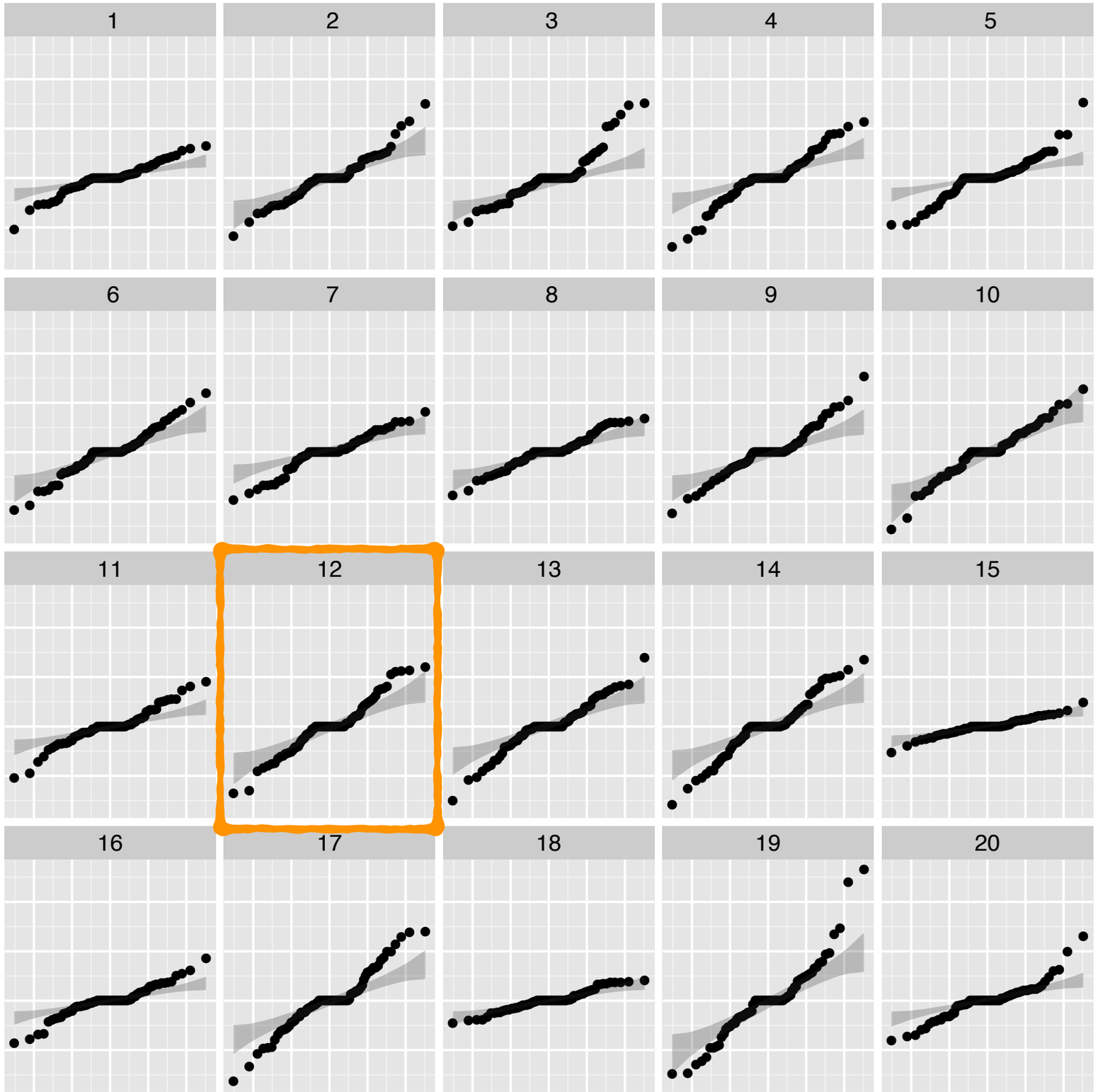
Predicted random effects for a HLM fit to radon in houses study. Nulls simulated from normal model assuming independence from the error terms.

Pick the plot that is most different from the others.



Predicted random effects for a HLM fit to radon in houses study. Nulls simulated from normal model assuming independence from the error terms.

Pick the plot that is most different from the others.



Predicted random effects for a HLM fit to radon in houses study. Nulls simulated from normal model assuming independence from the error terms.

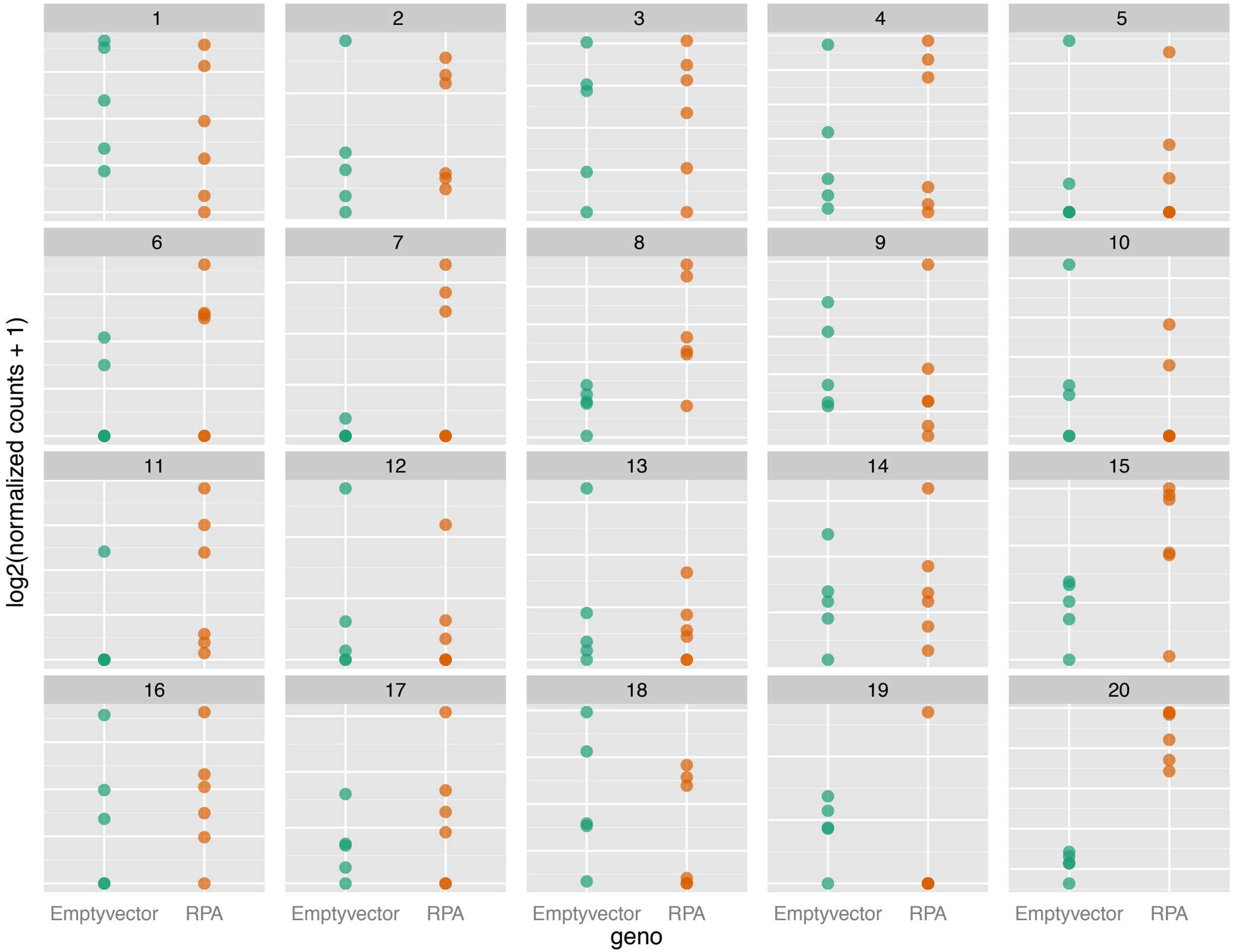
Pick the plot that is most different from the others.

16 of the 19 null plots fail the Anderson-Darling test of normality

Is there any structure in my RNA-seq data?

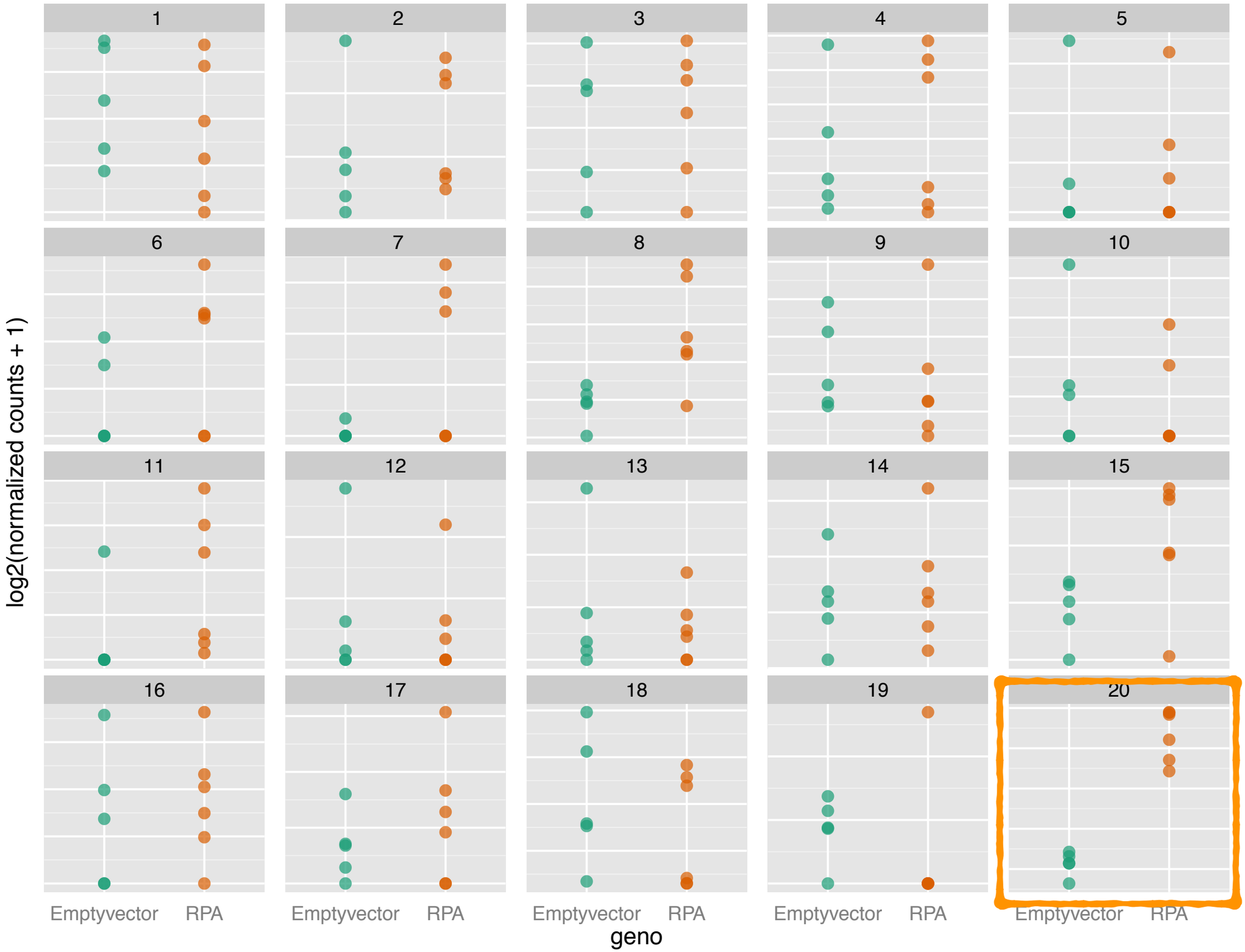
- Compare the results with random results
- Take the experimental design, 2x2x3, and permute the labels
- Re-run the analysis, record most significant gene
- Plot the results

In which of these plots do the two groups have the most vertical difference?



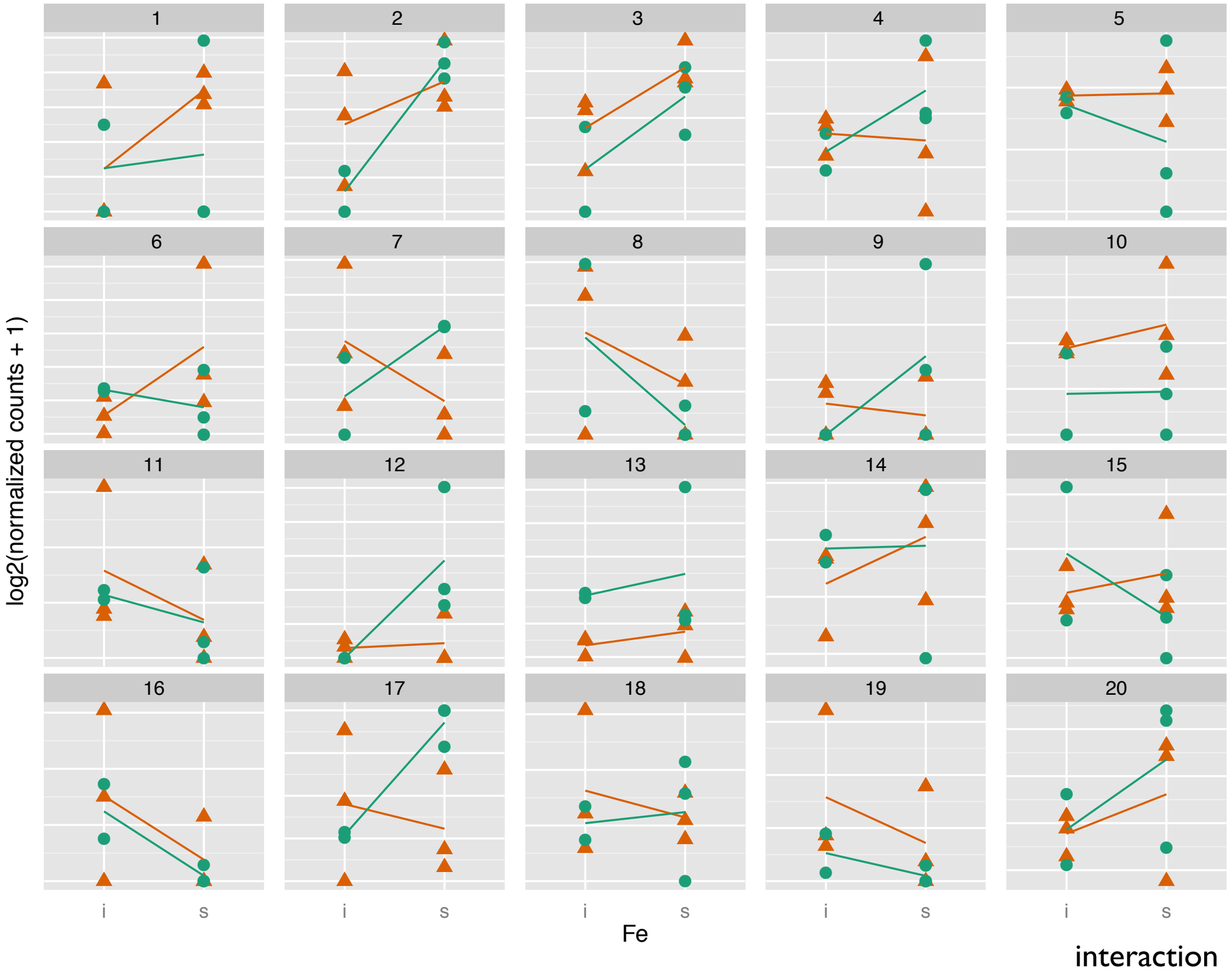
geno_1_5, 5/7

In which of these plots do the two groups have the most vertical difference?



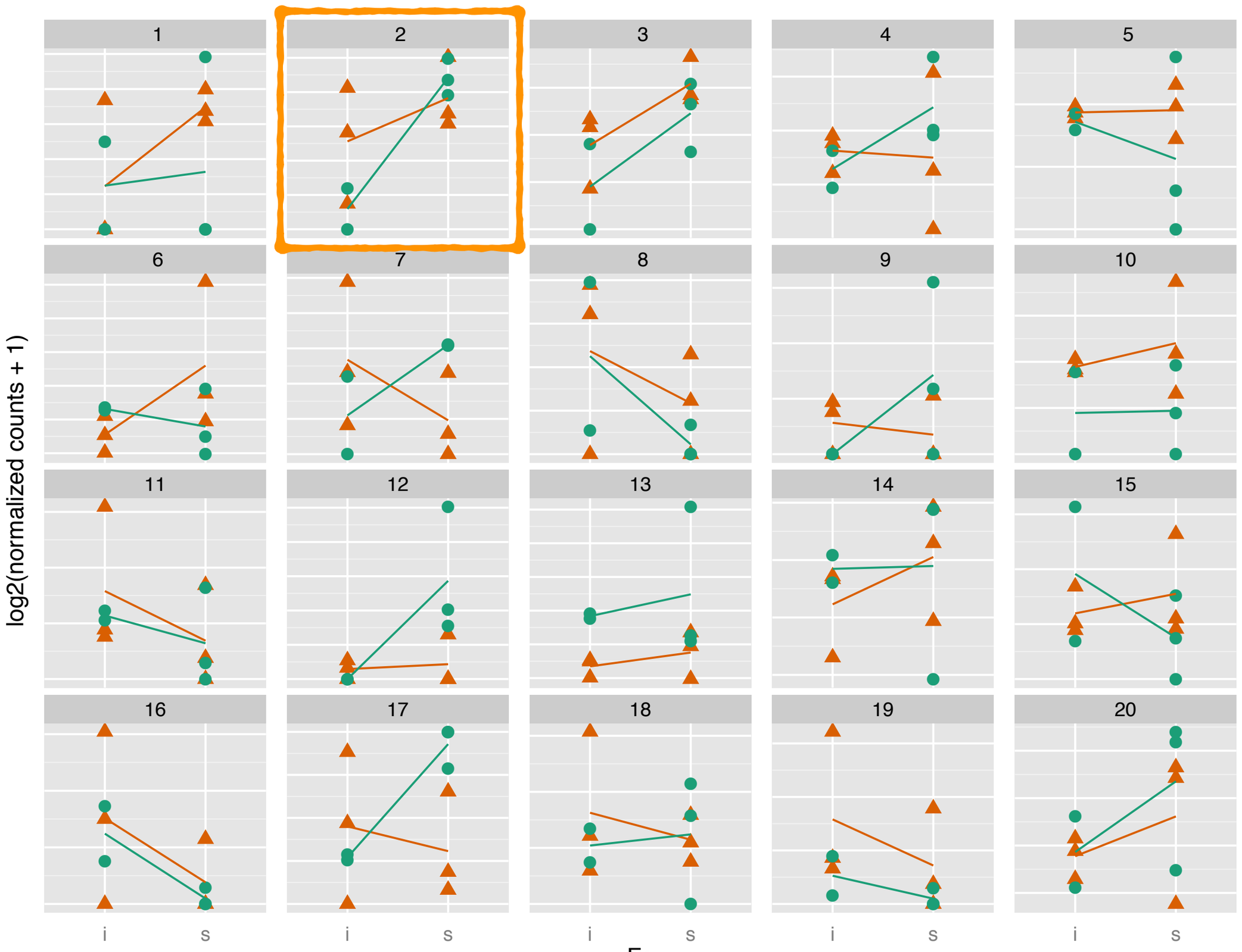
geno_1_5, 5/7

In which of these plots is the green line the steepest, and the spread of the green points relatively small?



interaction_2_1, 4/5

In which of these plots is the green line the steepest, and the spread of the green points relatively small?



interaction_2_1, 4/5

Experiment

- Five different sets of null plots
- Five different locations of true data plot inside the lineup
- Shown to a sample of Amazon Turk workers
- Overwhelmingly in both cases, the true data is picked, slightly less so for interaction

Experiment

- Five different sets of null plots
- Five different locations of true data plot inside the lineup
- Shown to a sample of Amazon Turk workers
- Overwhelmingly in both cases, the true data is picked, slightly less so for interaction

Data has SOME SIGNAL!

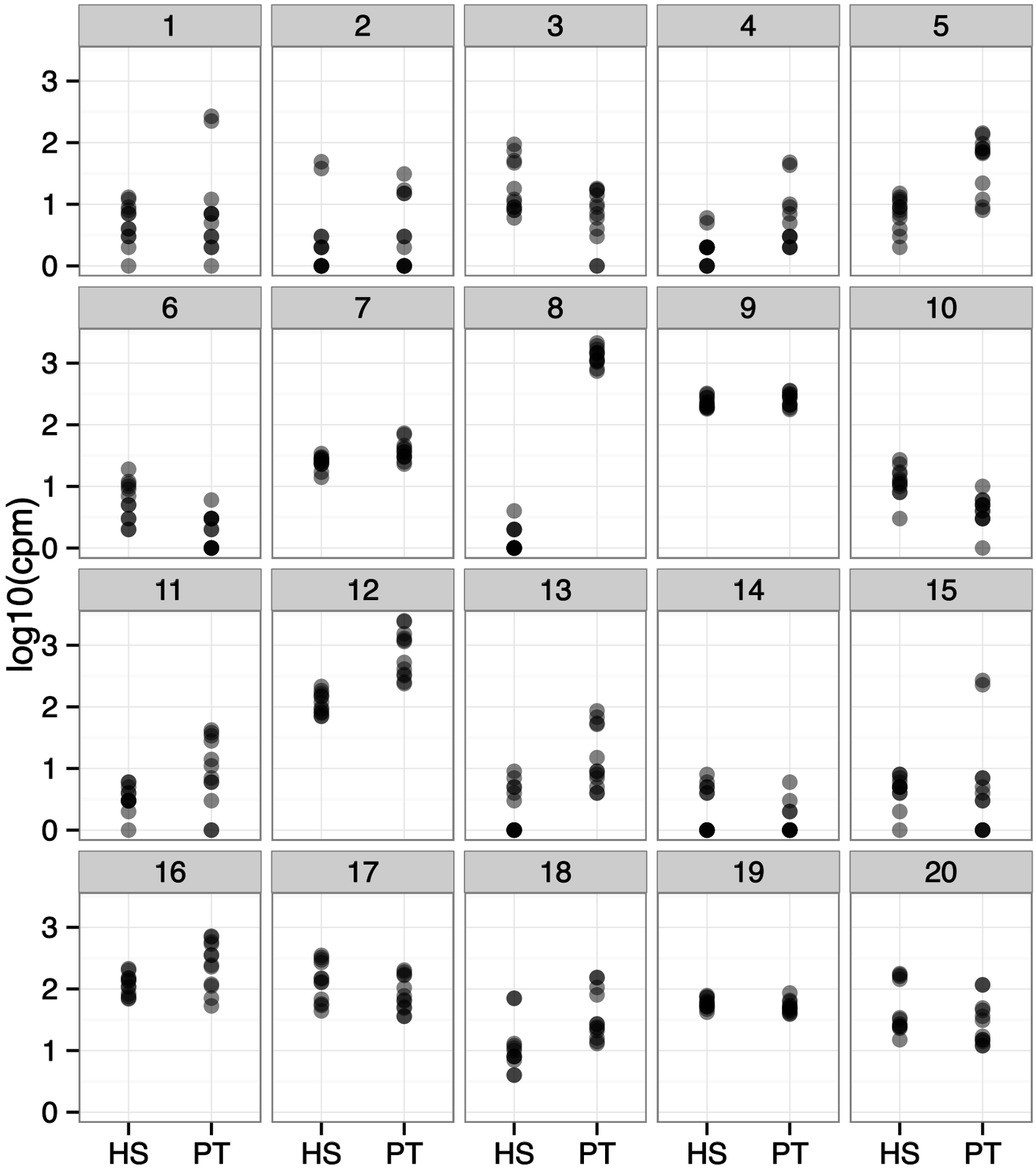
Data pulled from study

"Sex-specific and lineage-specific alternative splicing in primates" Blekhman, Marioni, Zumbo, Stephens, Gilad, Genome Research, 2010 20: 180-189,
http://genome.cshlp.org/content/suppl/2009/12/16/gr.099226.109.DC1.html

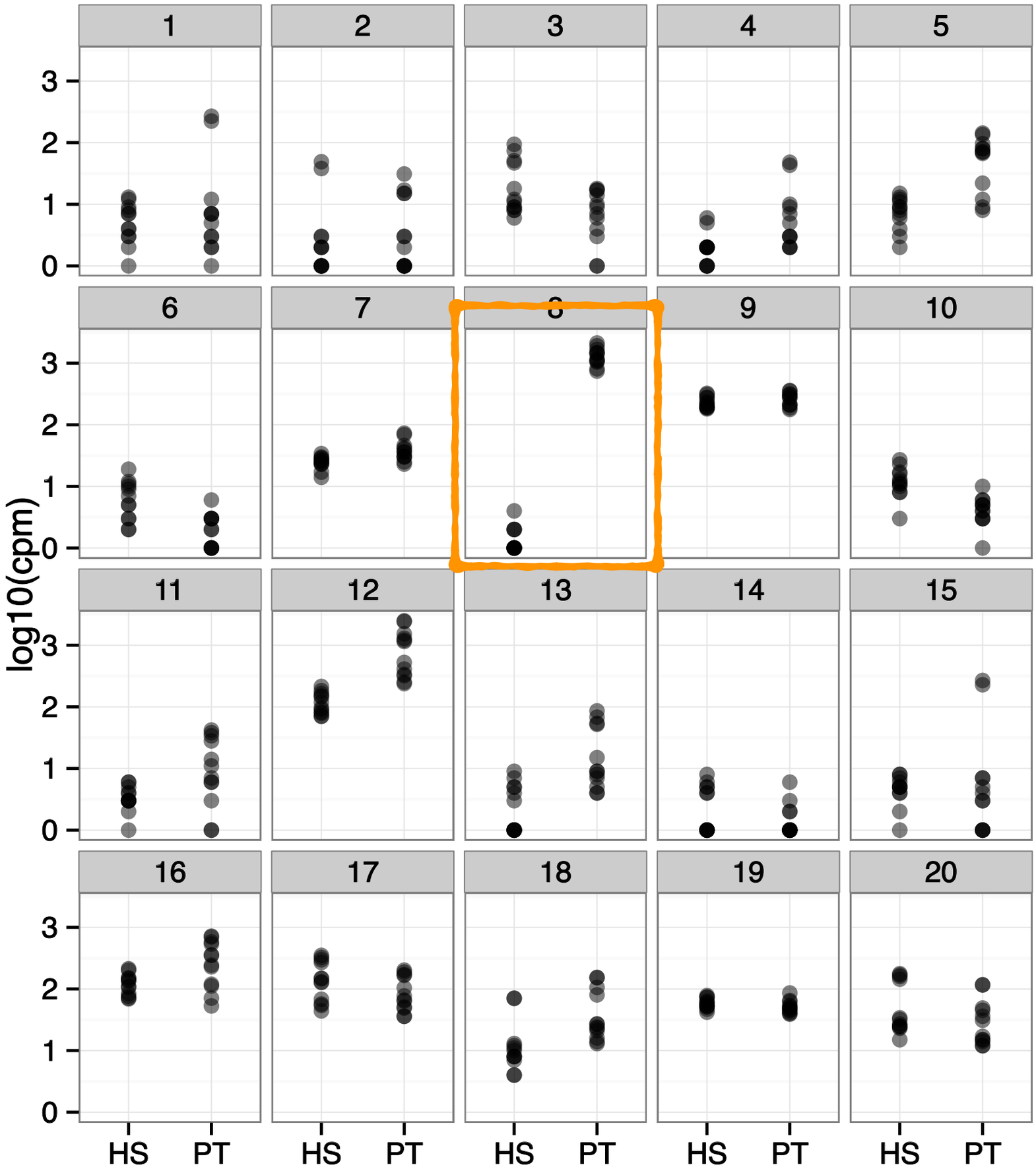
Claimed ~3000 significantly expressed genes

We permuted experimental design and re-ran 19 times pulled off top 3000 genes for each.

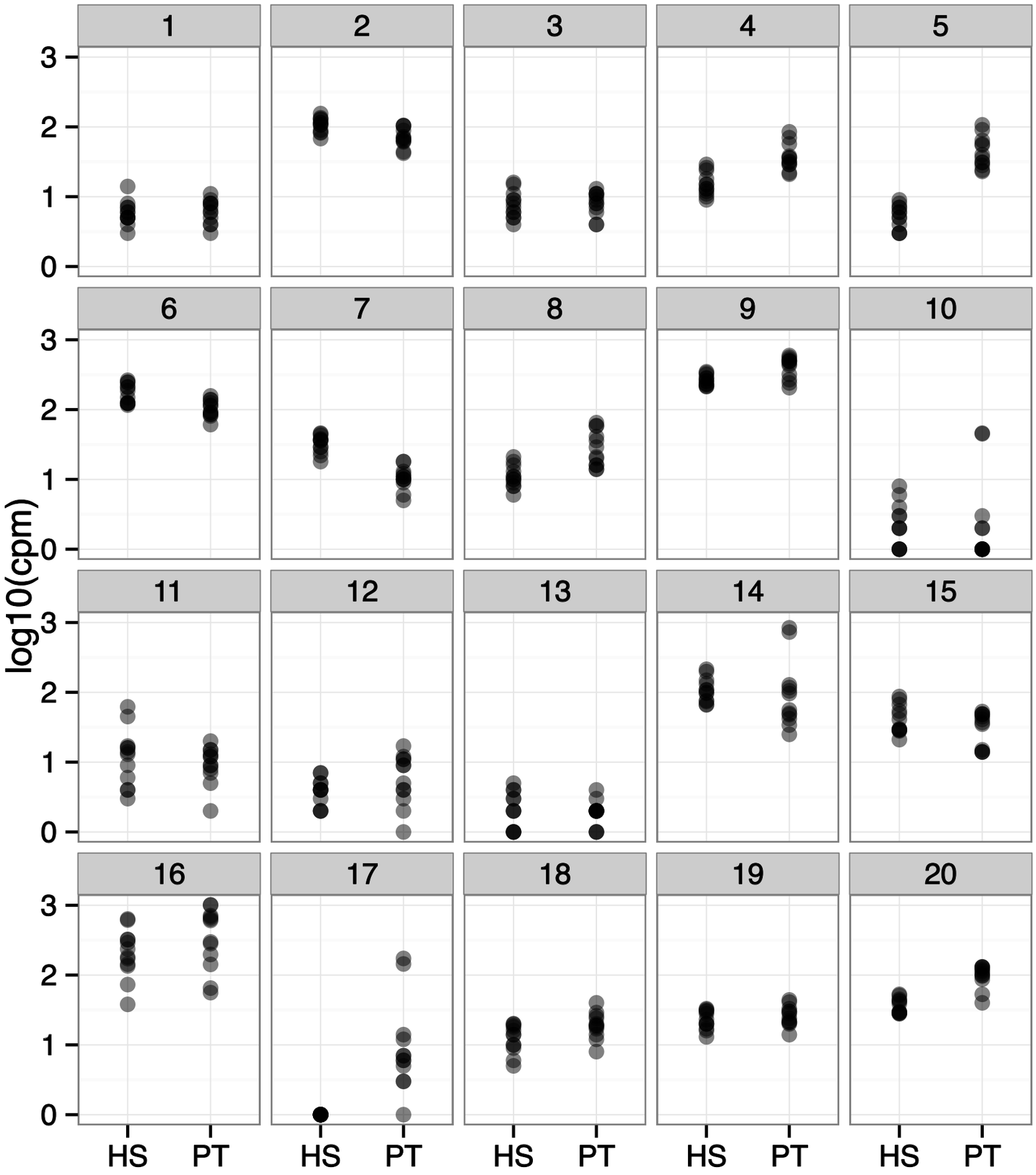
Human-chimp 1



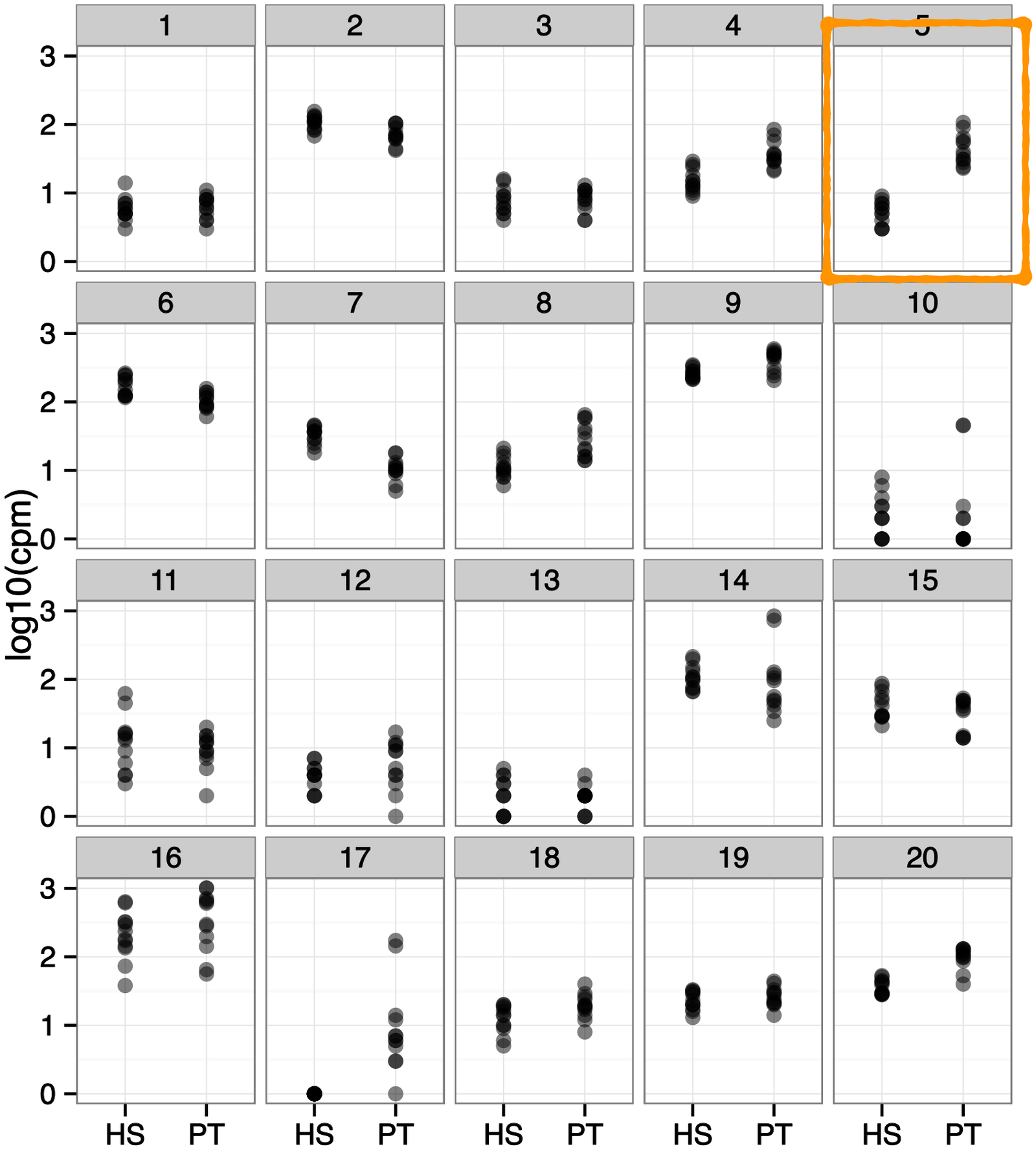
Human-chimp 1



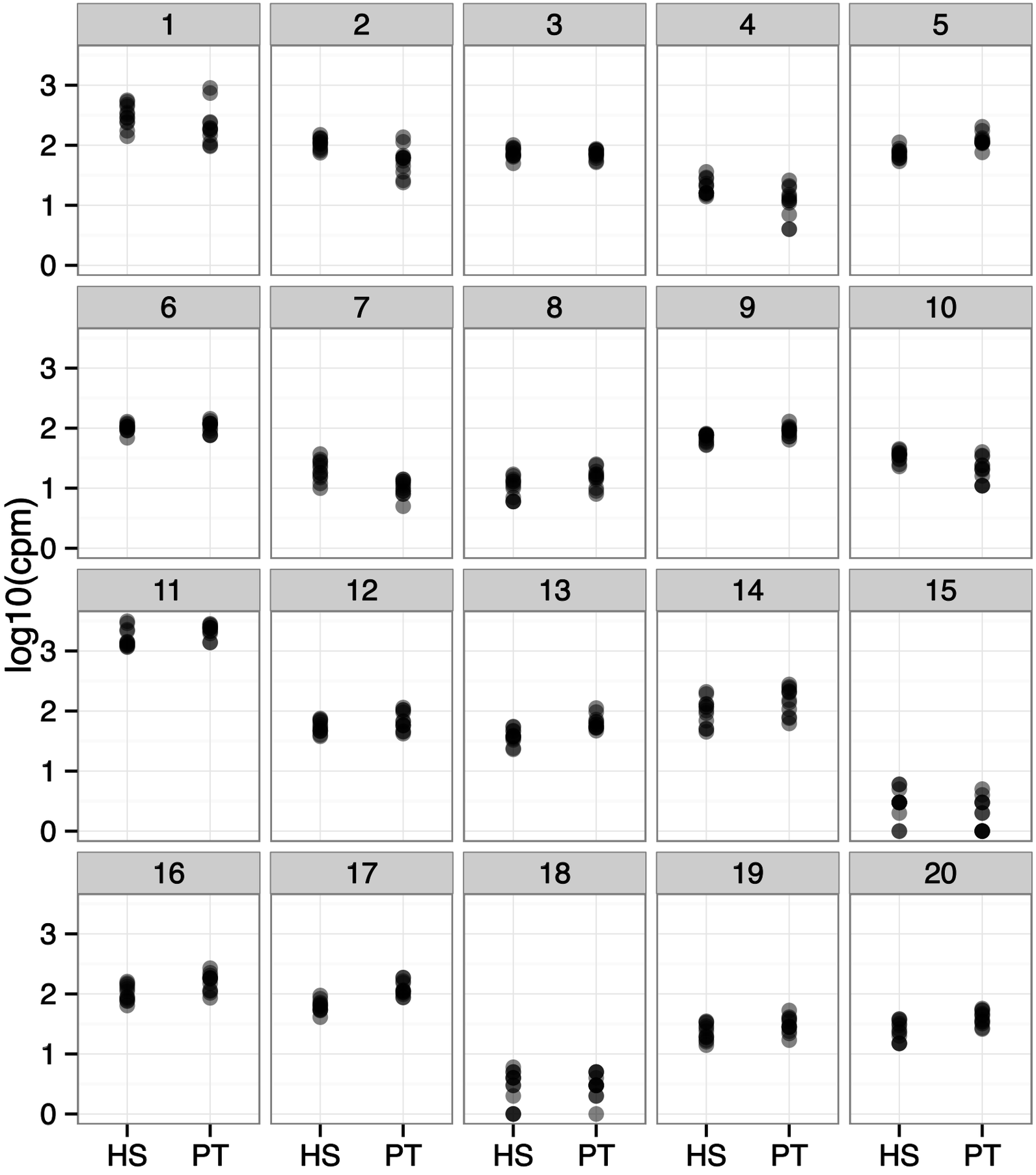
Human-chimp 2



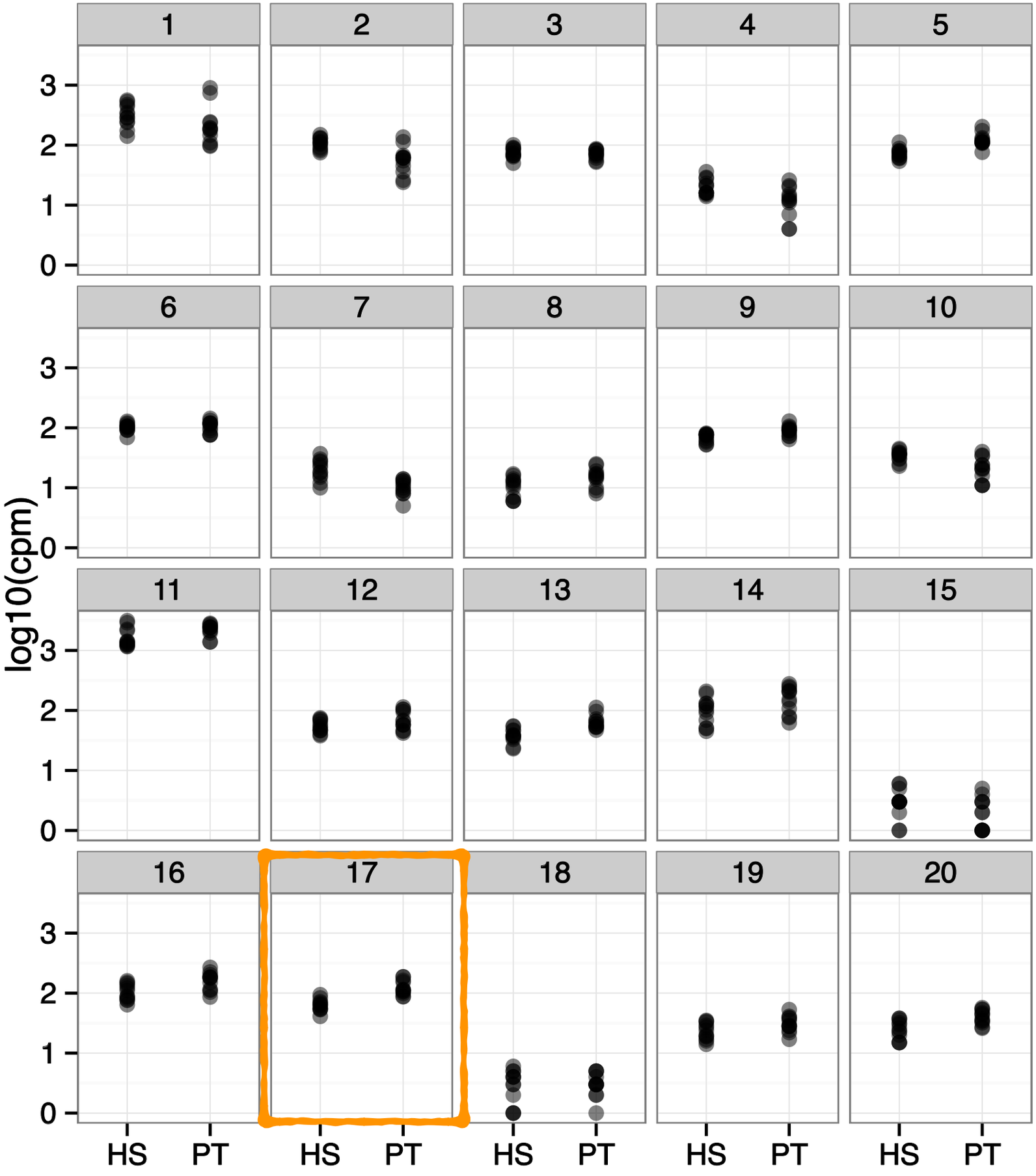
Human-chimp 2



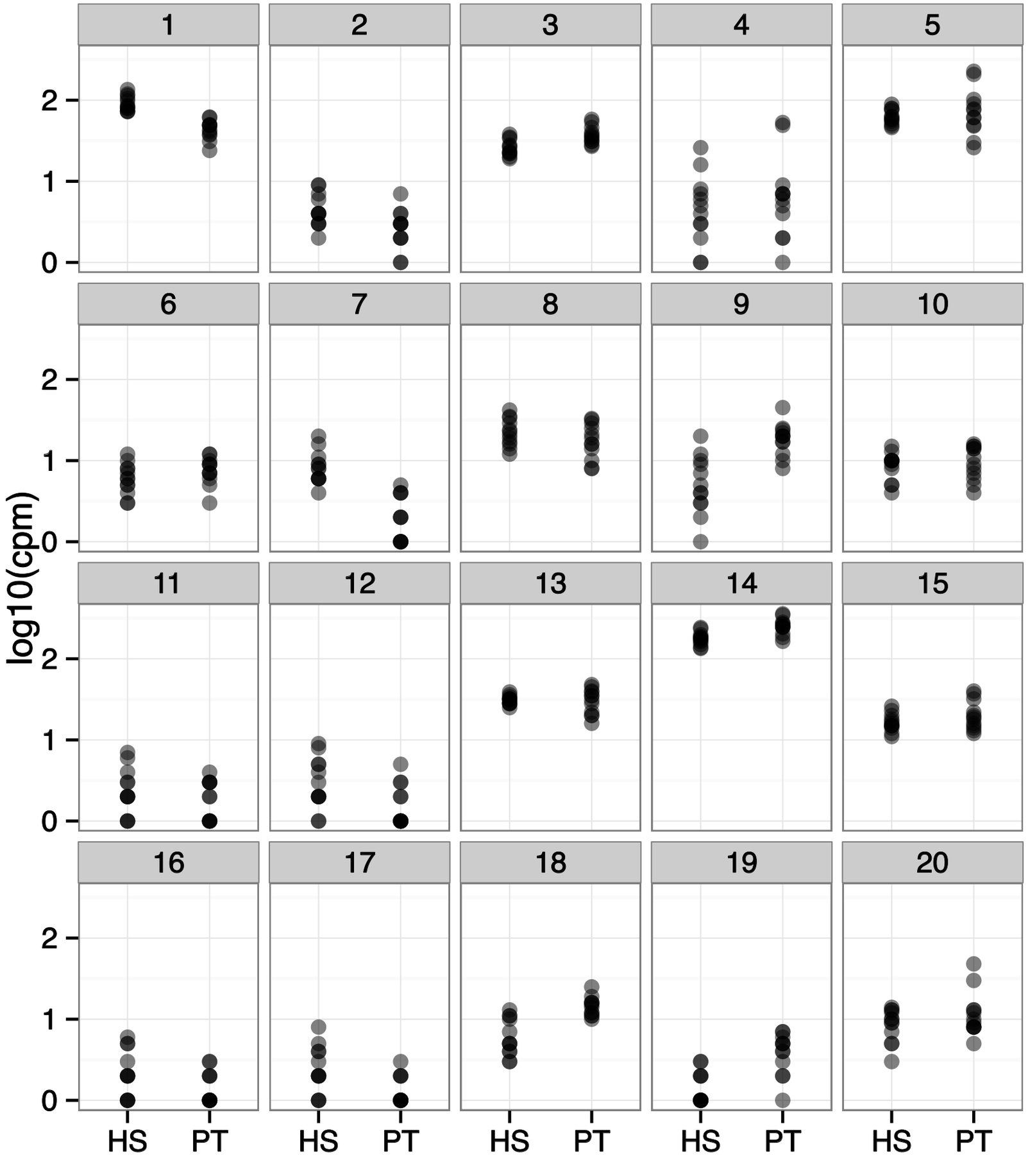
Human-chimp 3



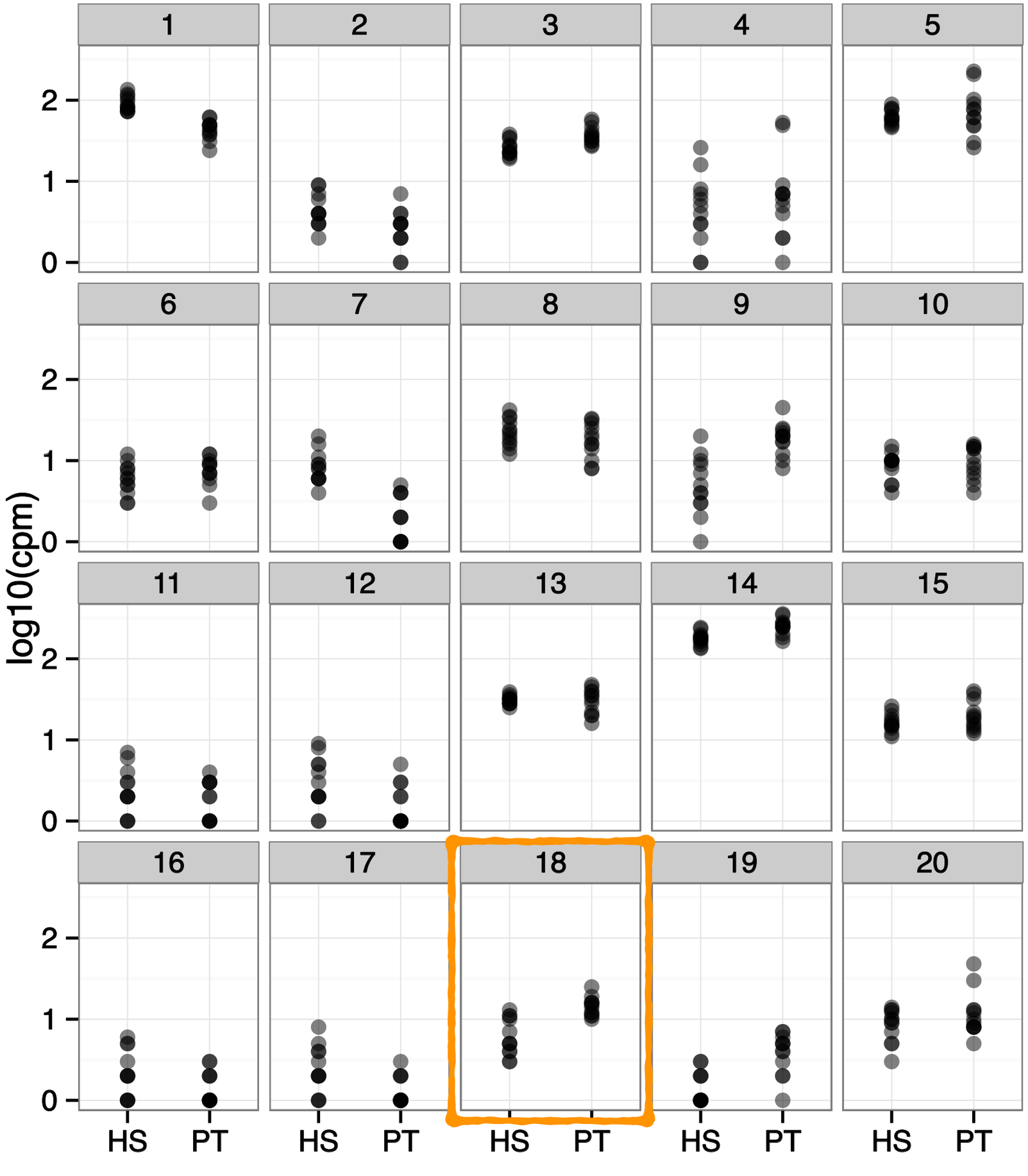
Human-chimp 3



Human-chimp 4



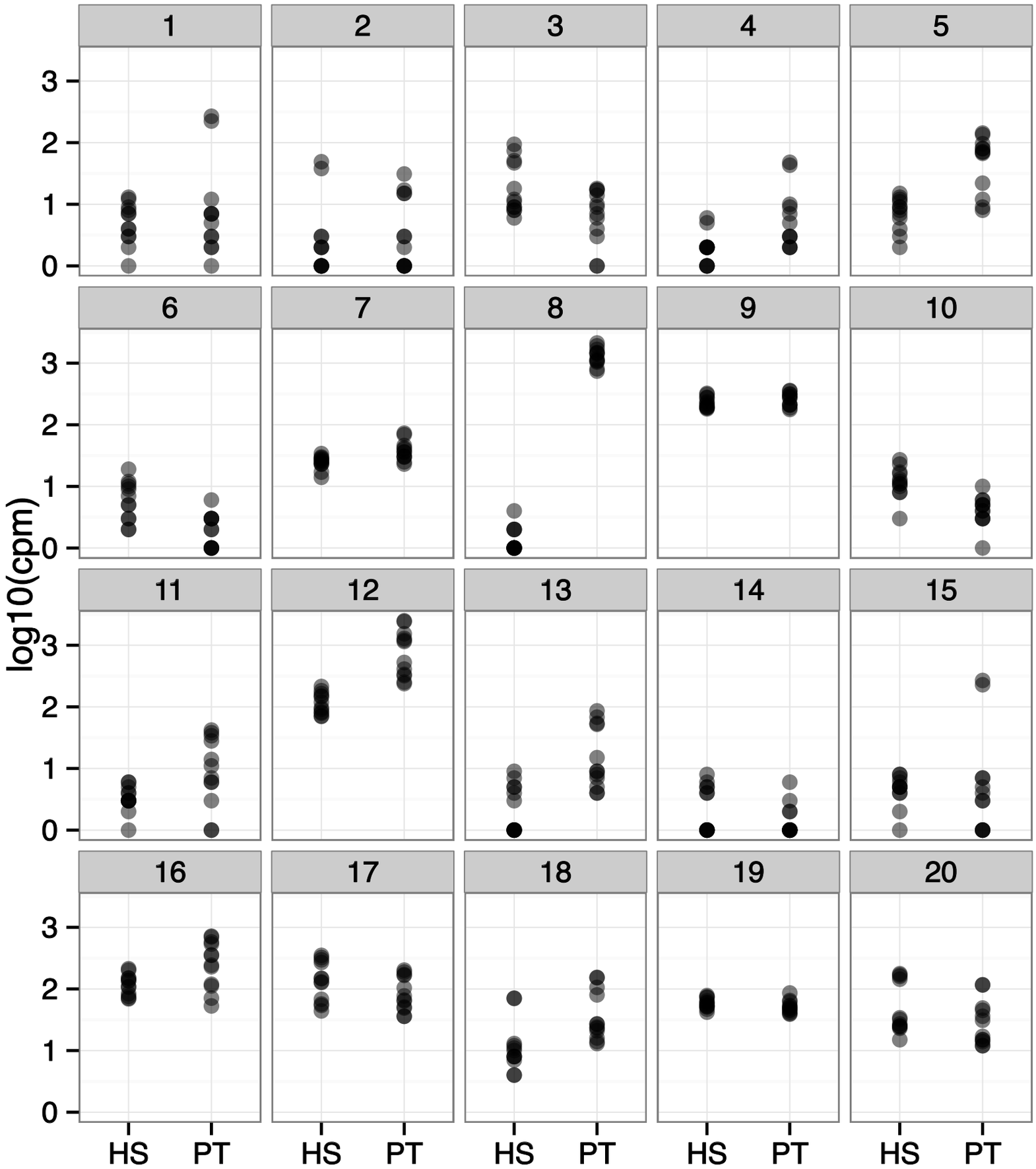
Human-chimp 4



What the Turk study showed

Closer to 100 significantly expressed genes.

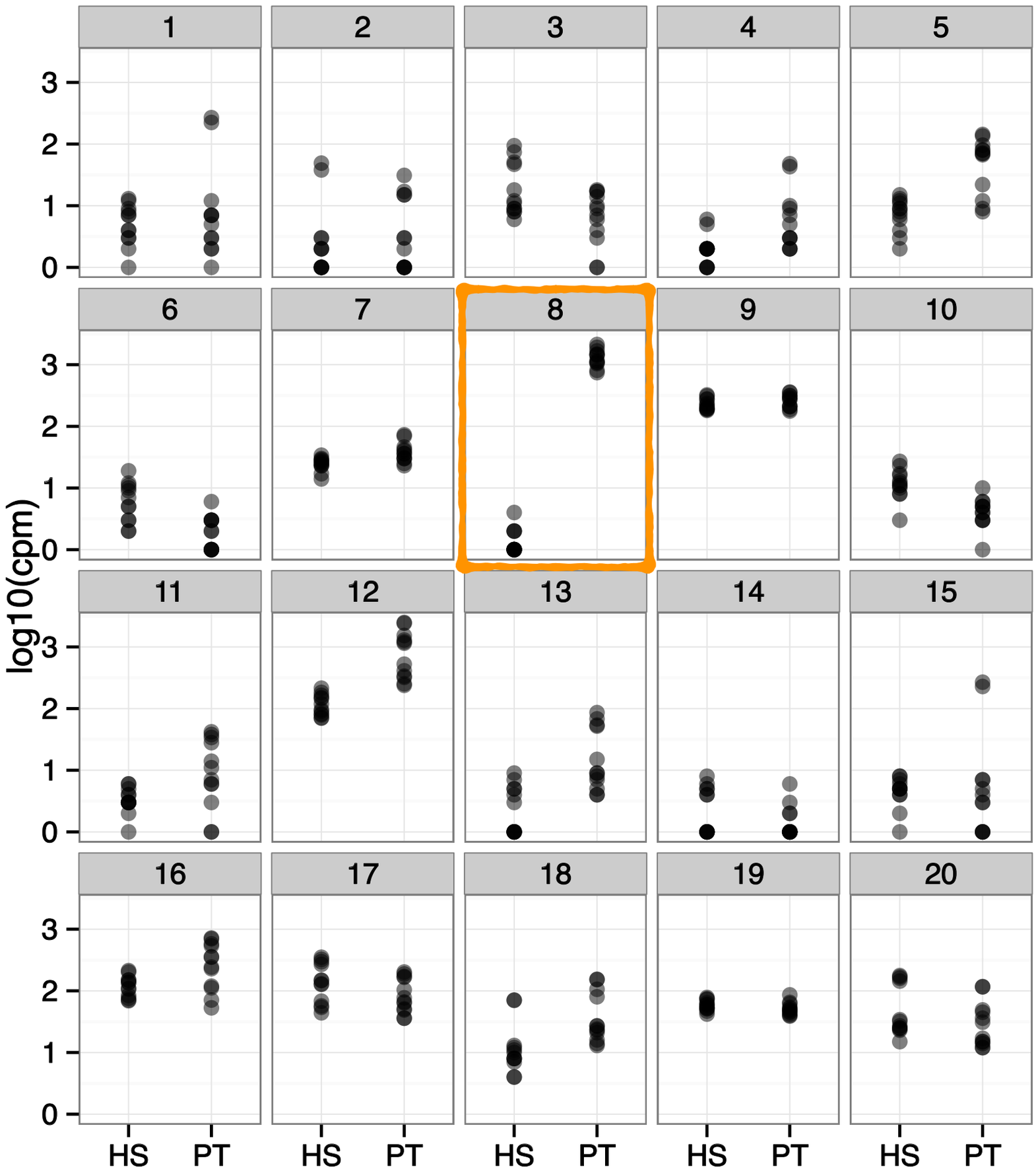
Human-chimp 1 2nd



Plot fold change against *p*-value to get at effect size - most interesting.

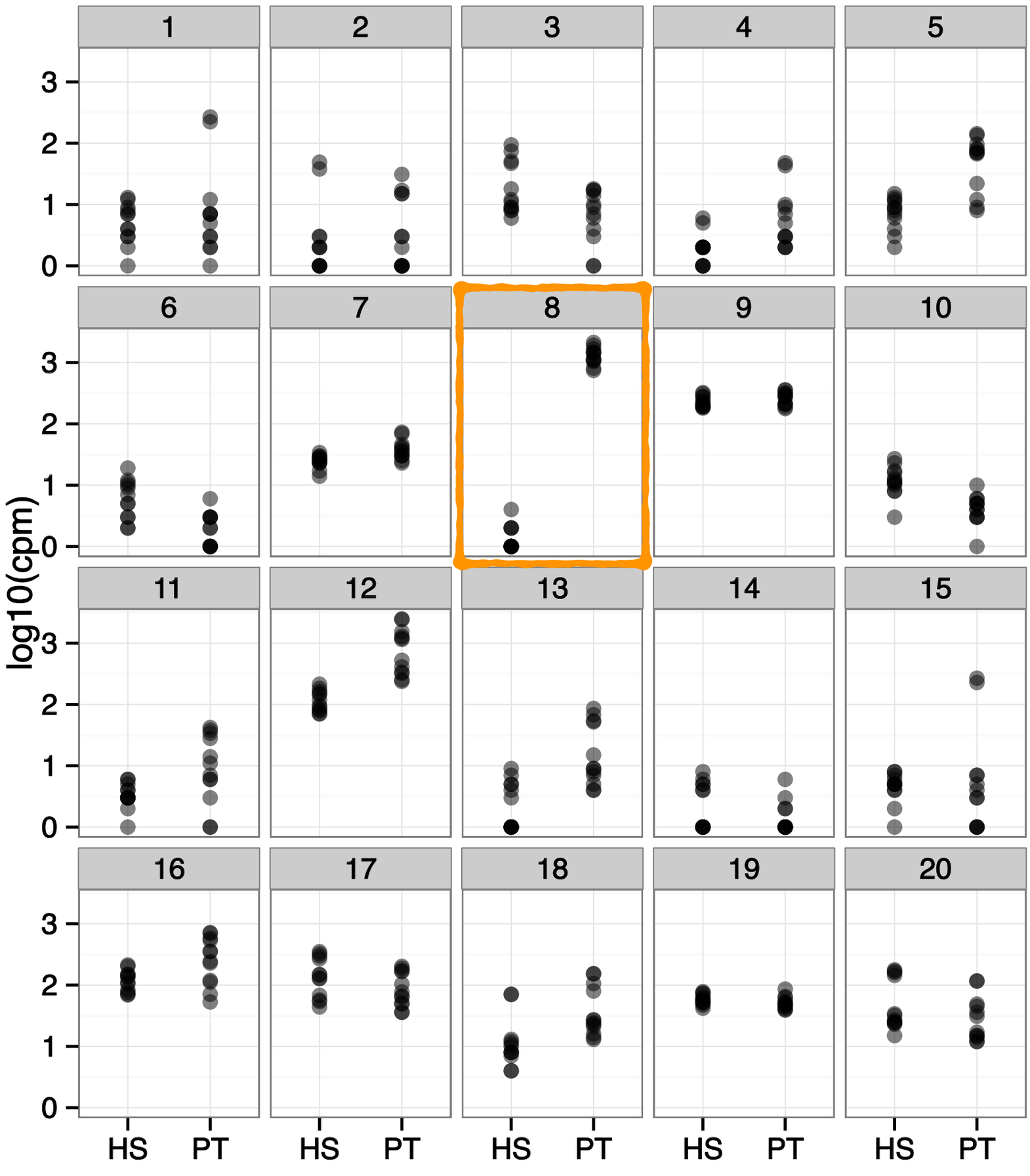
Human-chimp 1

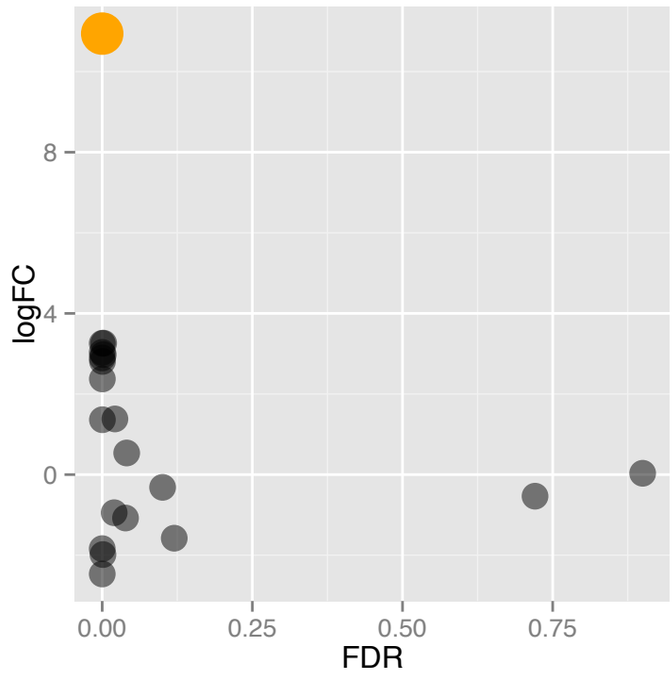
2nd



Plot fold change against *p*-value to get at effect size - most interesting.

Human-chimp 1 2nd

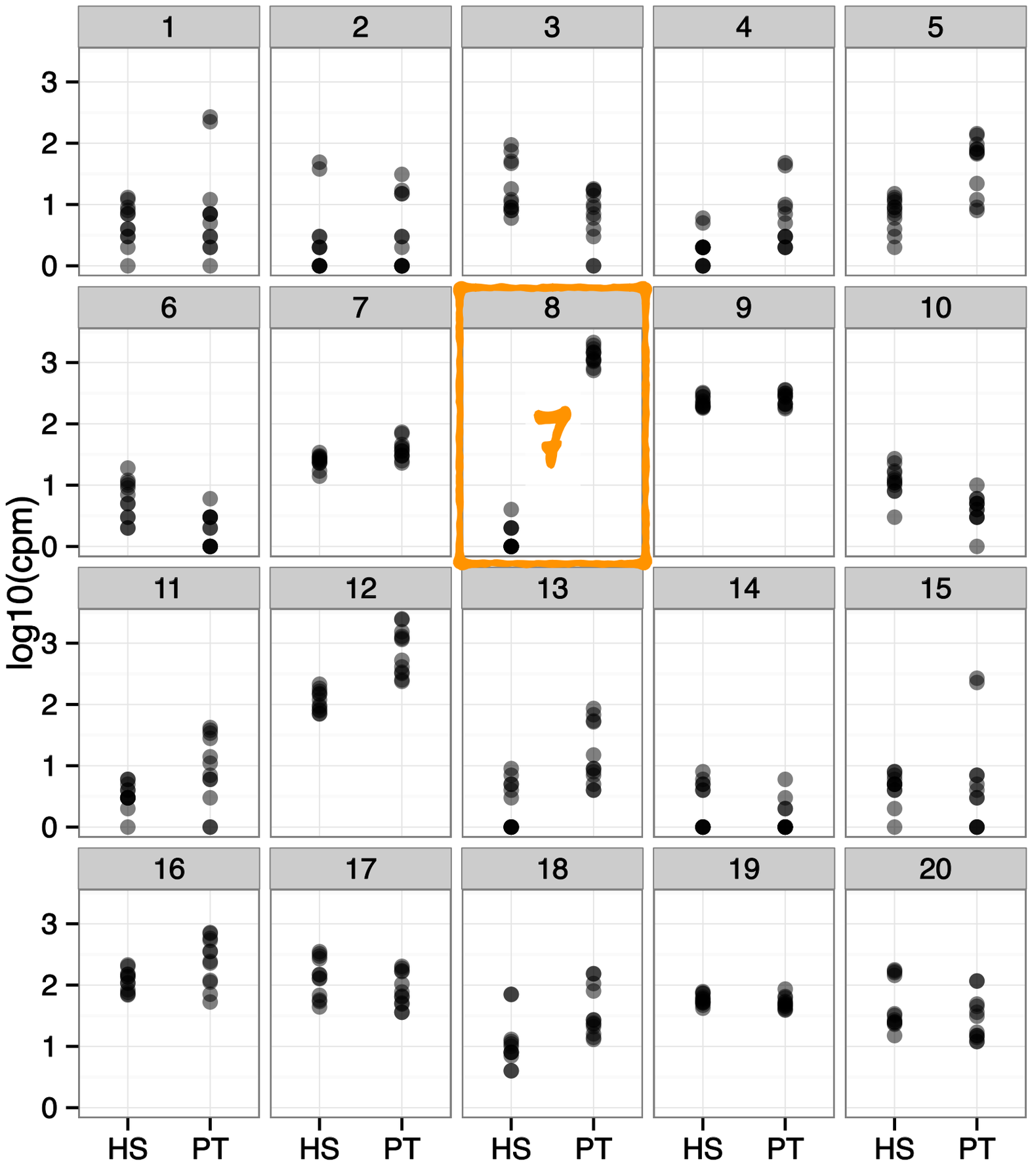


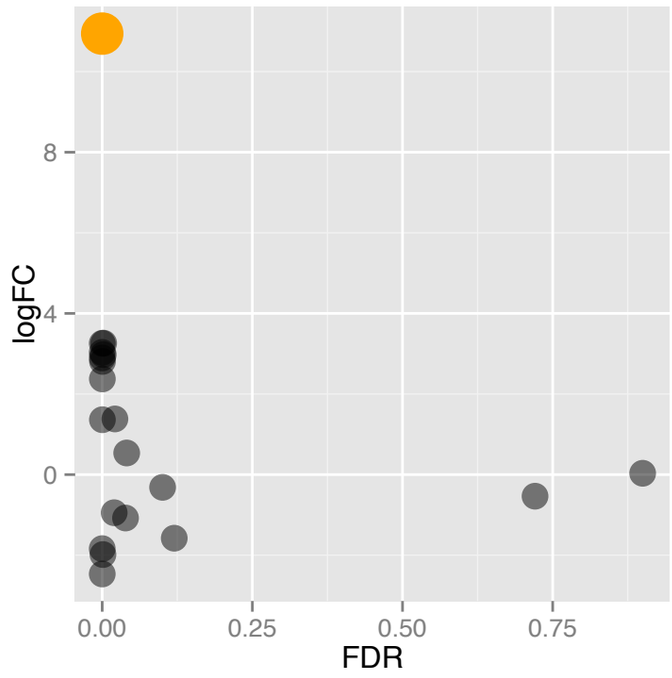


Plot fold change against *p*-value to get at effect size - most interesting.

Human-chimp 1

2nd

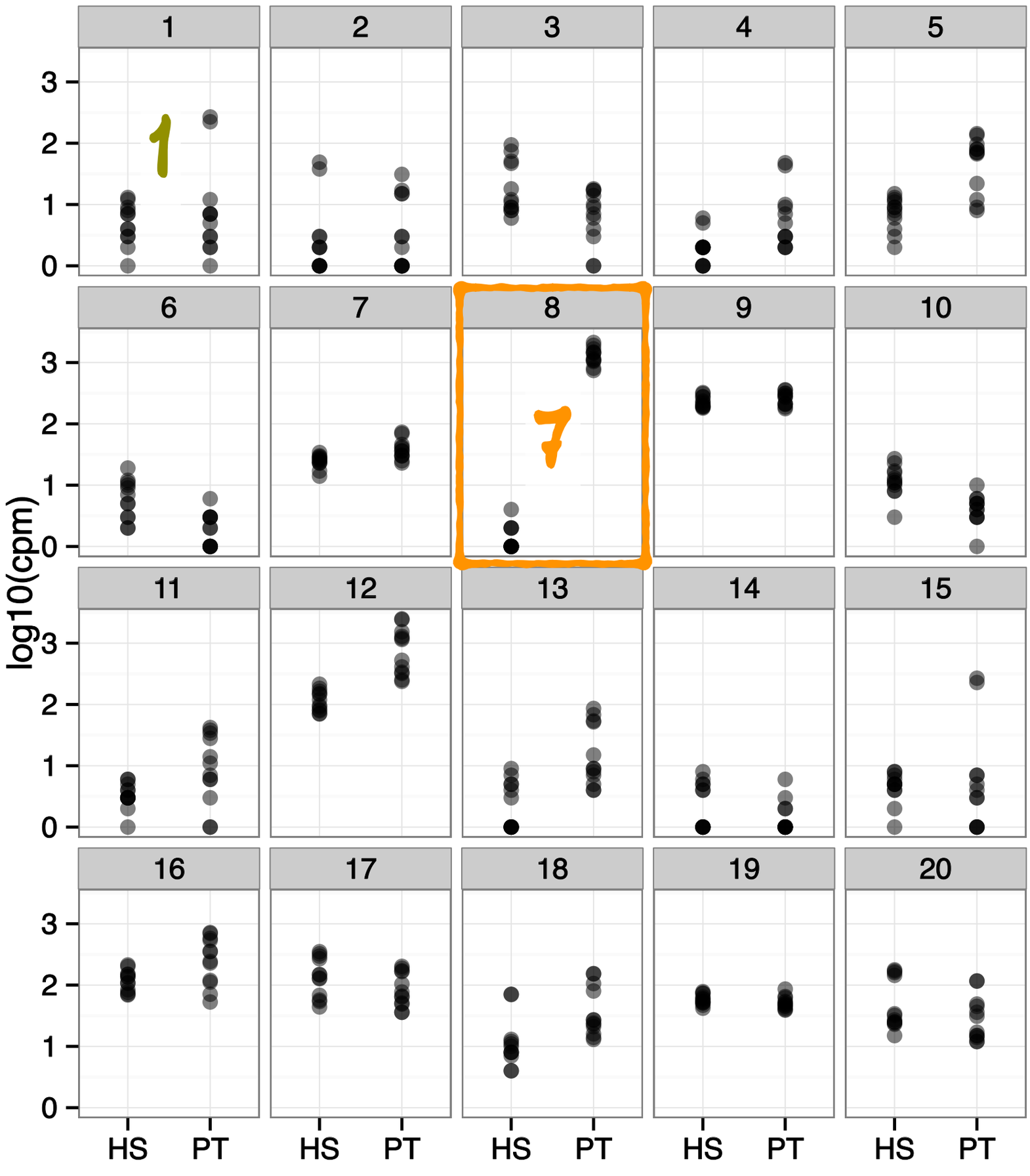


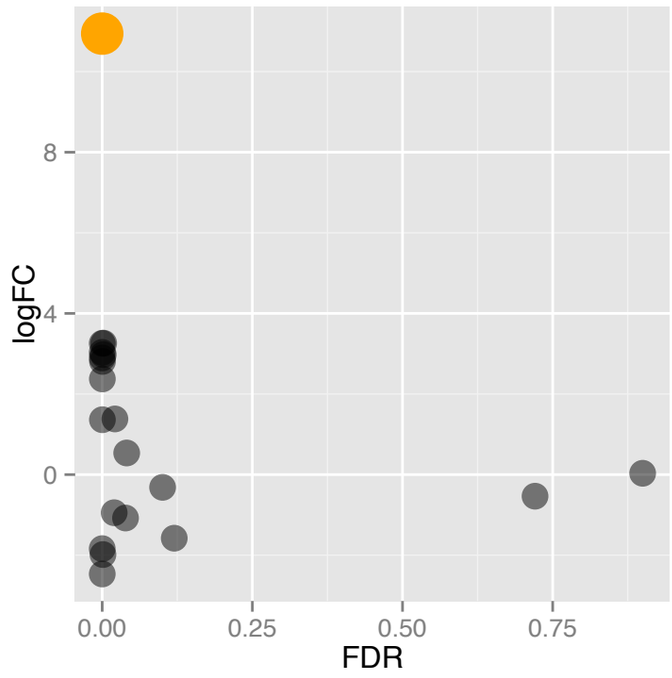


Plot fold change against *p*-value to get at effect size - most interesting.

Human-chimp 1

2nd

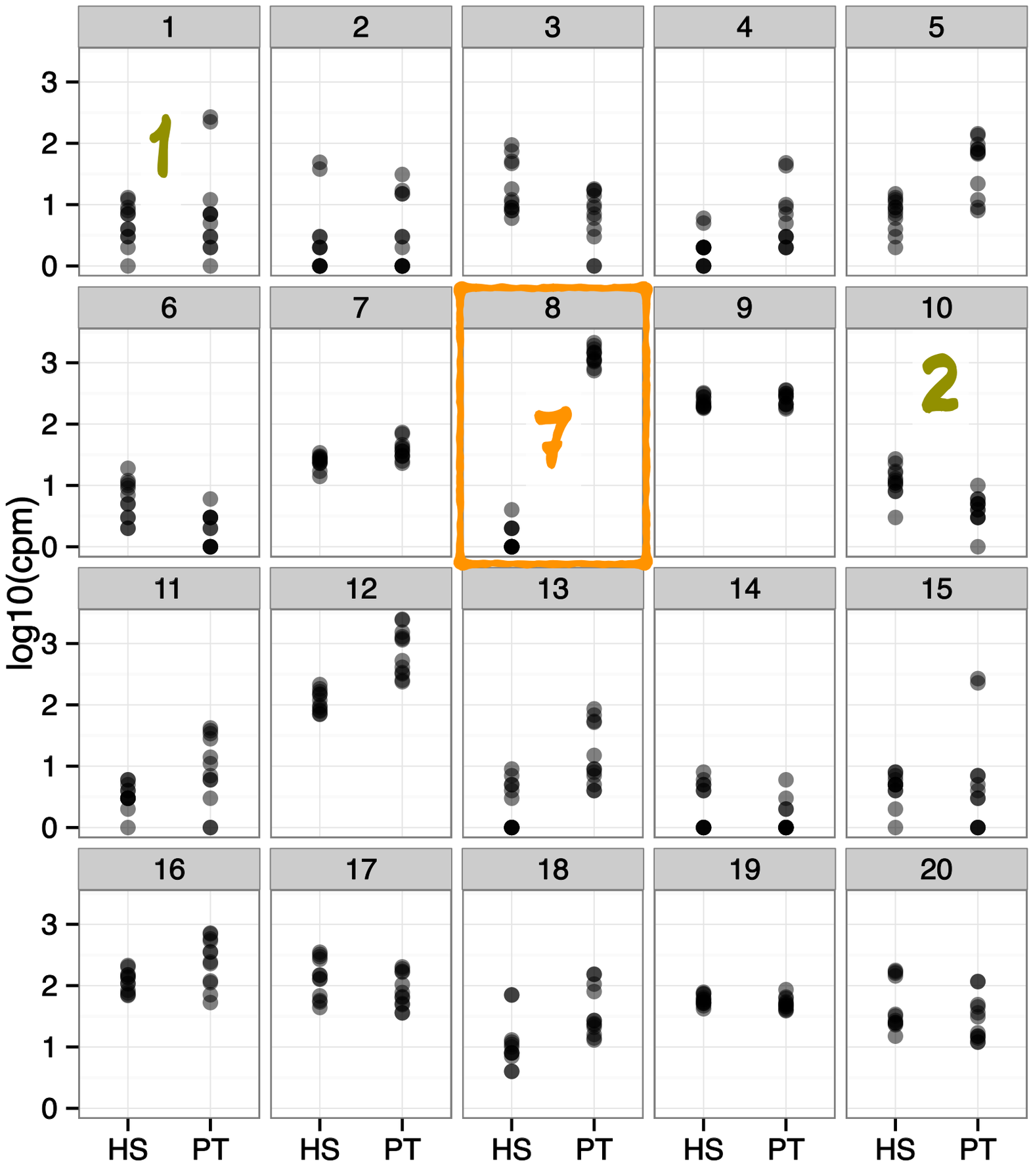


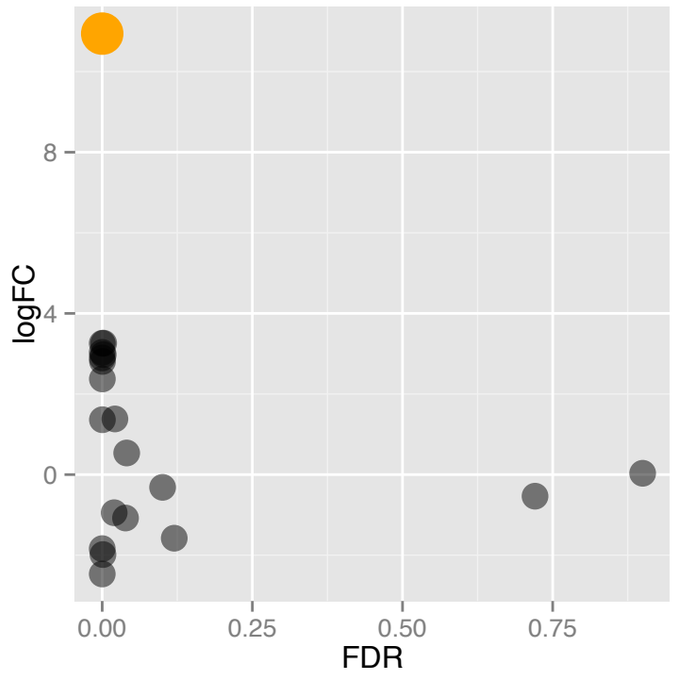


Plot fold change against *p*-value to get at effect size - most interesting.

Human-chimp 1

2nd

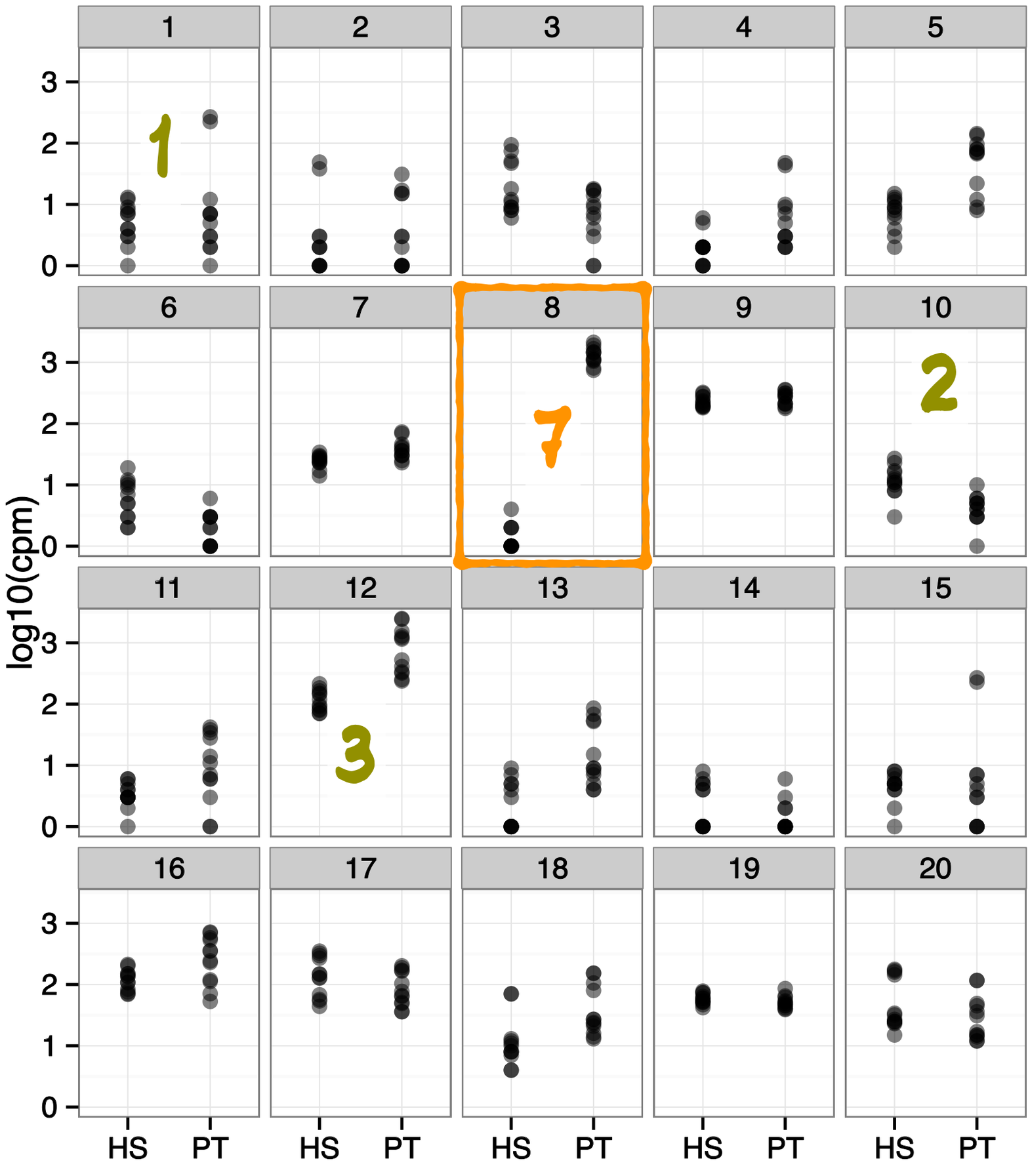


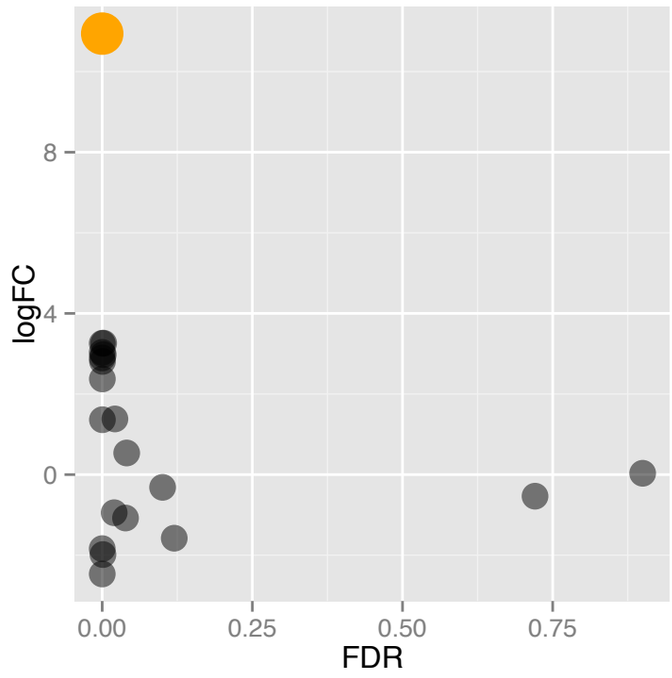


Plot fold change against *p*-value to get at effect size - most interesting.

Human-chimp 1

2nd

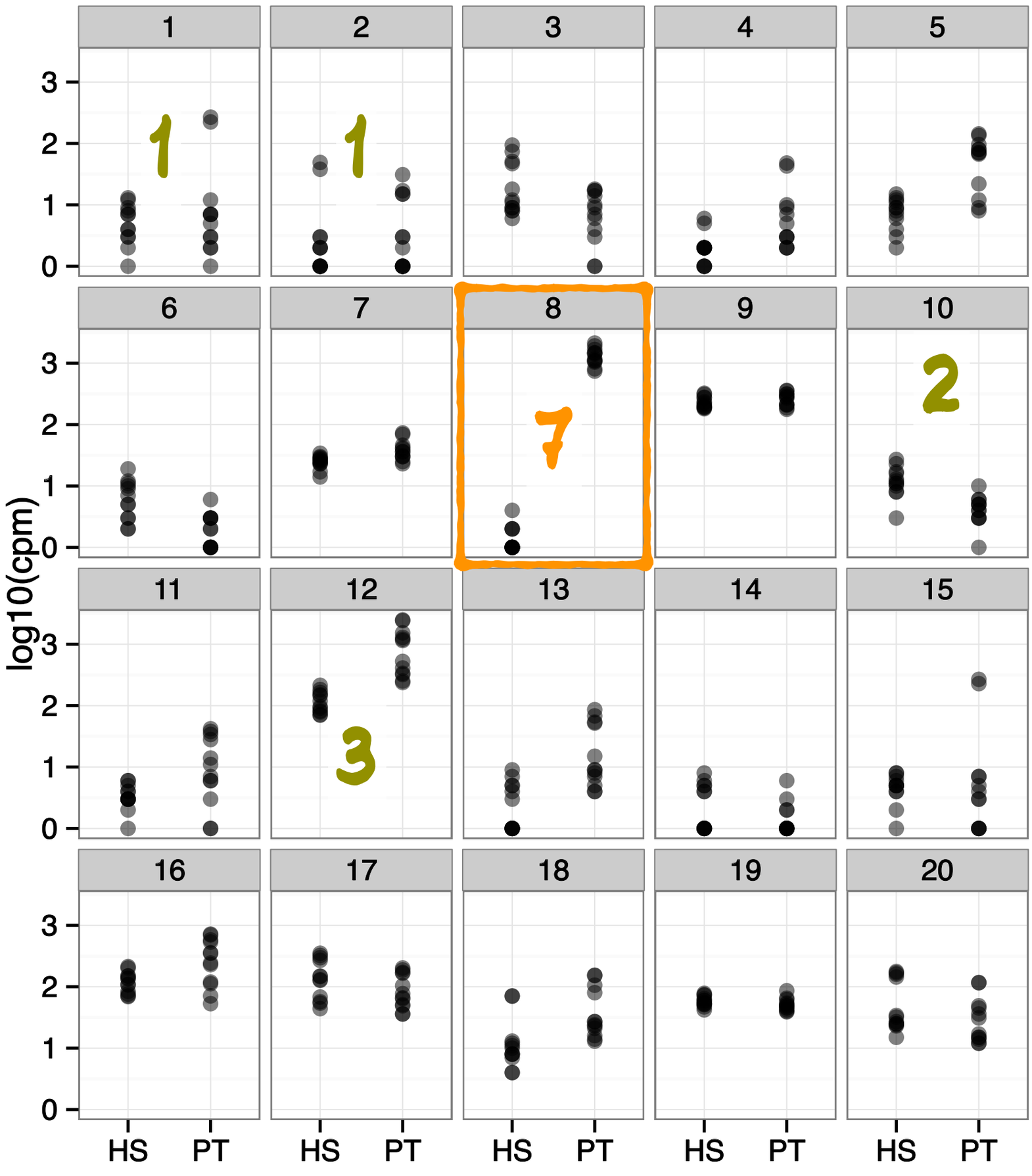


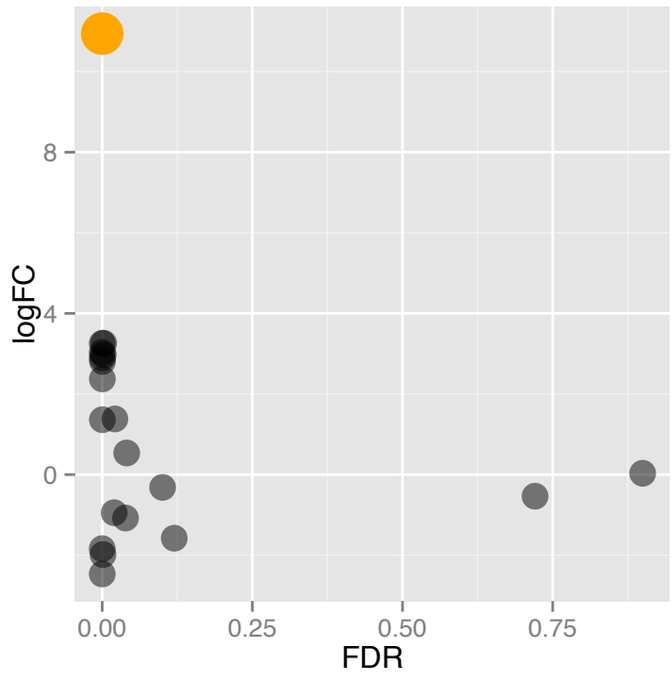


Plot fold change against *p*-value to get at effect size - most interesting.

Human-chimp 1

2nd





Plot fold change against *p*-value to get at effect size - most interesting.

Summary

- This is an exciting time to be a statistician, if you are interested in data analysis.
- Data is so widely available and accessible, that it is easy to extract, analyze and learn about our world.
- It is possible to both explore and yet maintain a healthy skepticism.

EDA & Inference

If the plot that is picked is the plot of the real data, this is statistical significance, and a *p*-value can be placed on the discovery.

- Buja et al (2009) RSPT A (econ eg)
- Wickham et al (2010) InfoVis/TVCG
- Hofmann et al (2012) InfoVis/TVCG
- Majumder et al (2013) JASA
- Roy Chowdhury et al (2014) Comput. Stat.
- Zhao et al (2014) IJITAS

Acknowledgements

- R packages: nullabor, ggplot2, XML, scrapeR
- Experiments web site: http://www.public.iastate.edu/~hofmann/experiments.html