Harnessing Crowd-Sourcing to Assess Genes based on Effect Size Using Visual Inference Methods

Di Cook, Monash University
Joint work with Niladri Roy Chowdhury, Eric Hare, Mahbub Majumder, Michelle Graham, Tengfei Yin, Heike Hofmann



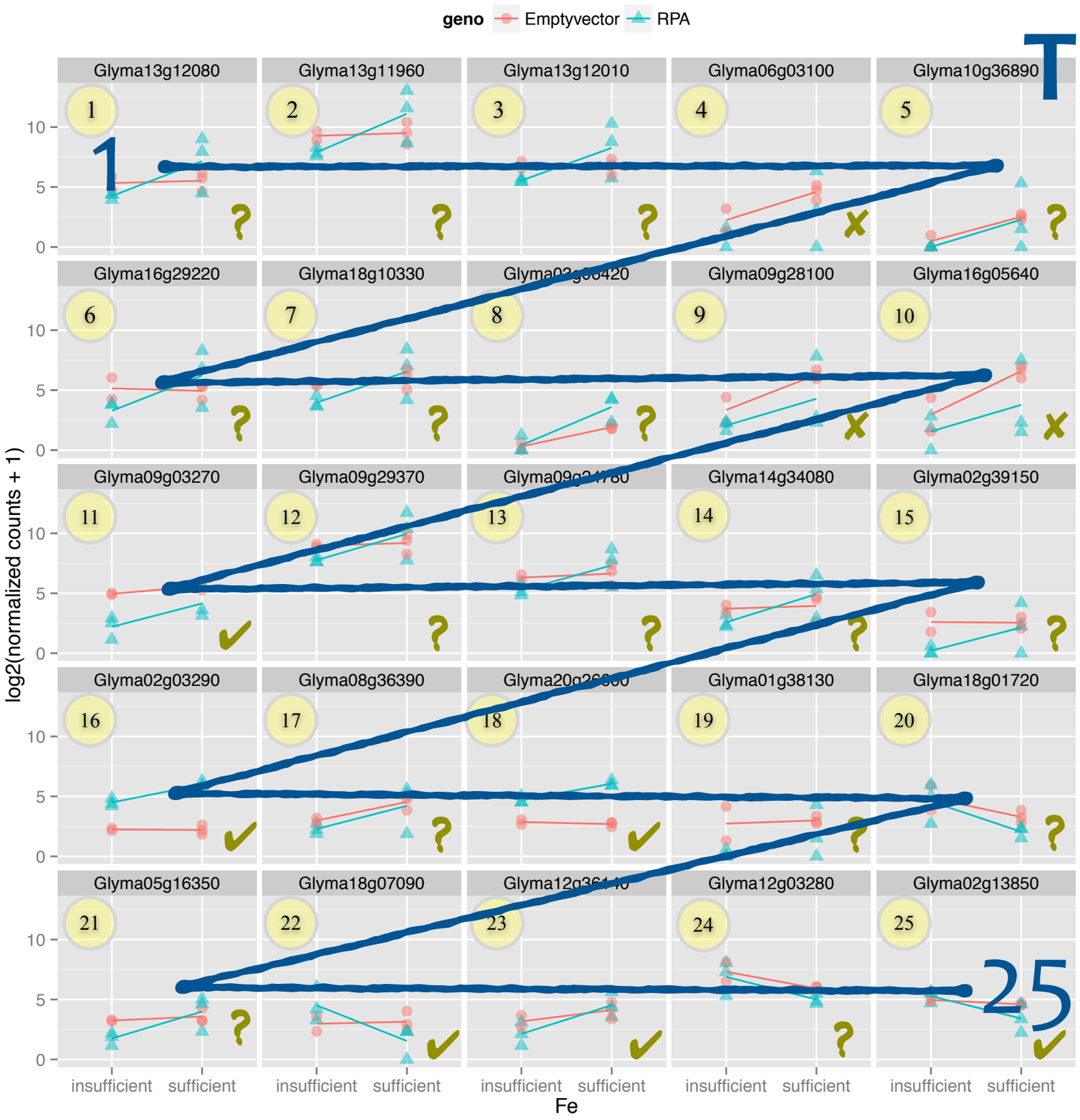



Outline

- Analysis outline, edgeR, … background
- Our top genes: good, maybe, ugly
- Why - video of dispersion
- First experiment, is there any structure
- Re-analysis of published study

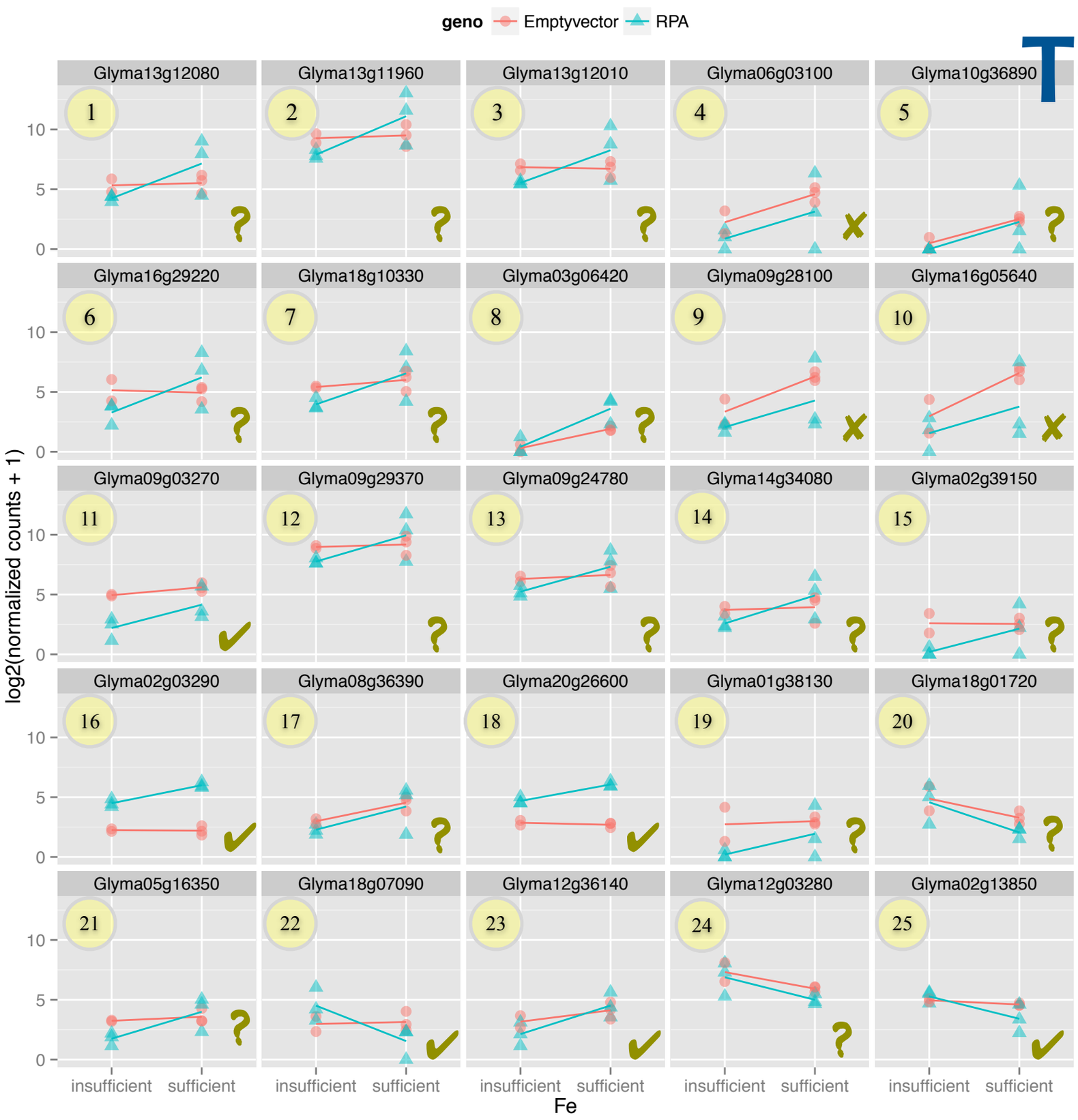
Our Data

- RNA libraries sequenced by Illumina HiSeq2000
- Alignment by `bowtie`
- `Rsamtools` to import bam files, `rtracklayer` to import gff files
- `GenomicRanges` to count reads
- Negative binomial model using `edgeR` to compute differential expression
- FDR yields ~2000 significantly expressed genes



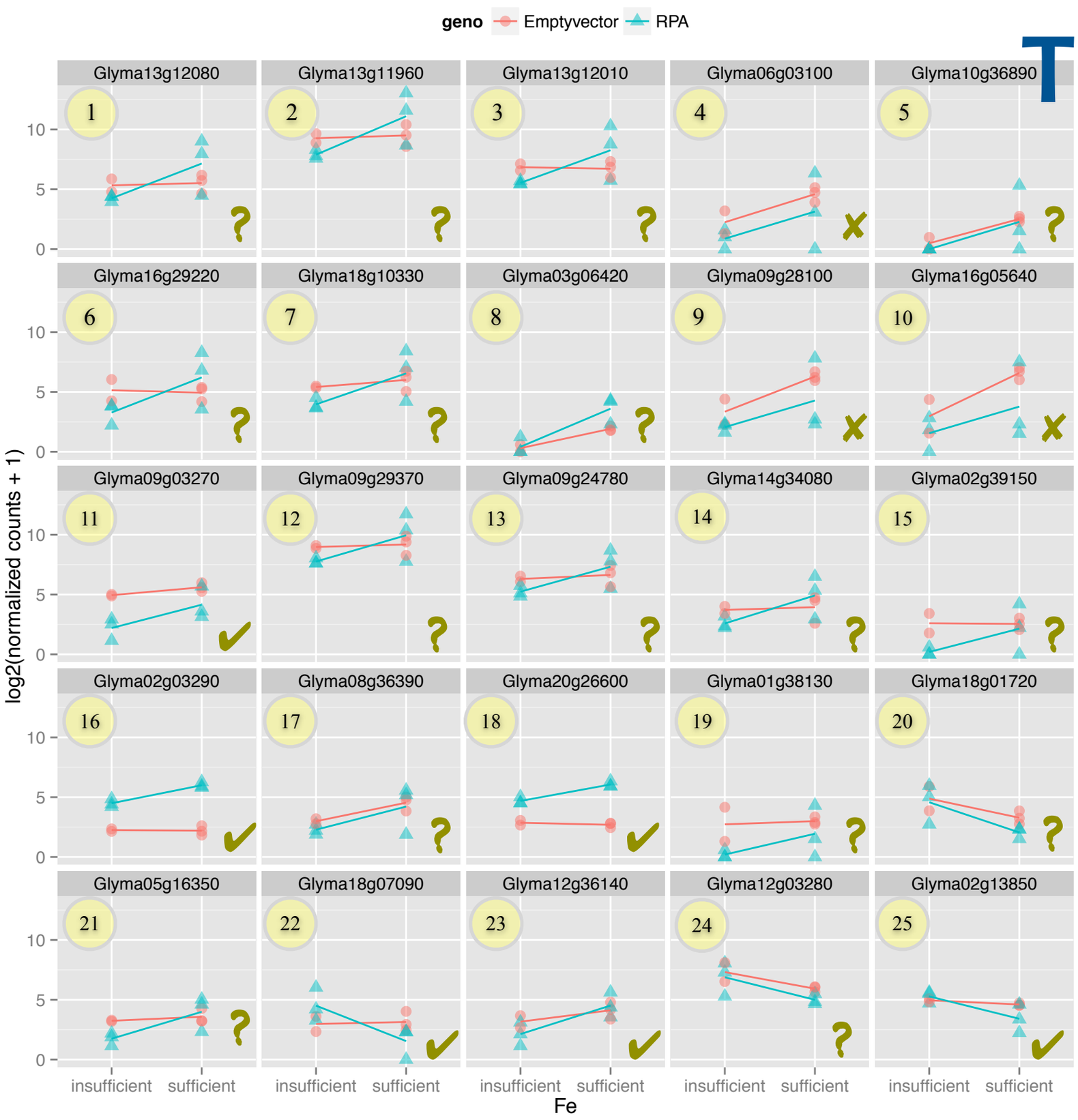
TOP 25 GENES

The Good (✔), Maybe (?) & Ugly (✘) ordered list of genes



TOP 25 GENES

The Good (✔), Maybe (?) & Ugly (✘) ordered list of genes



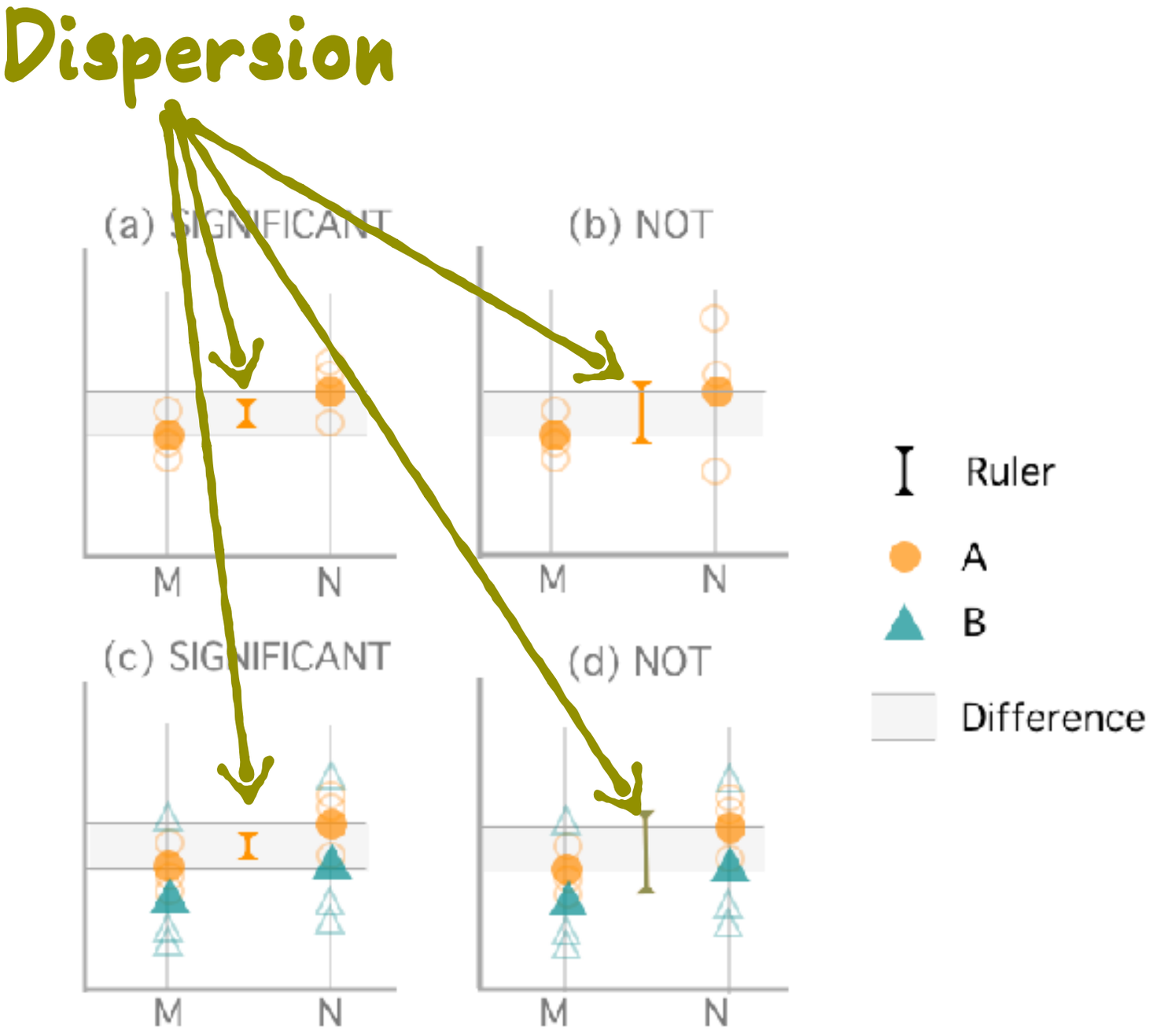
TOP 25 GENES

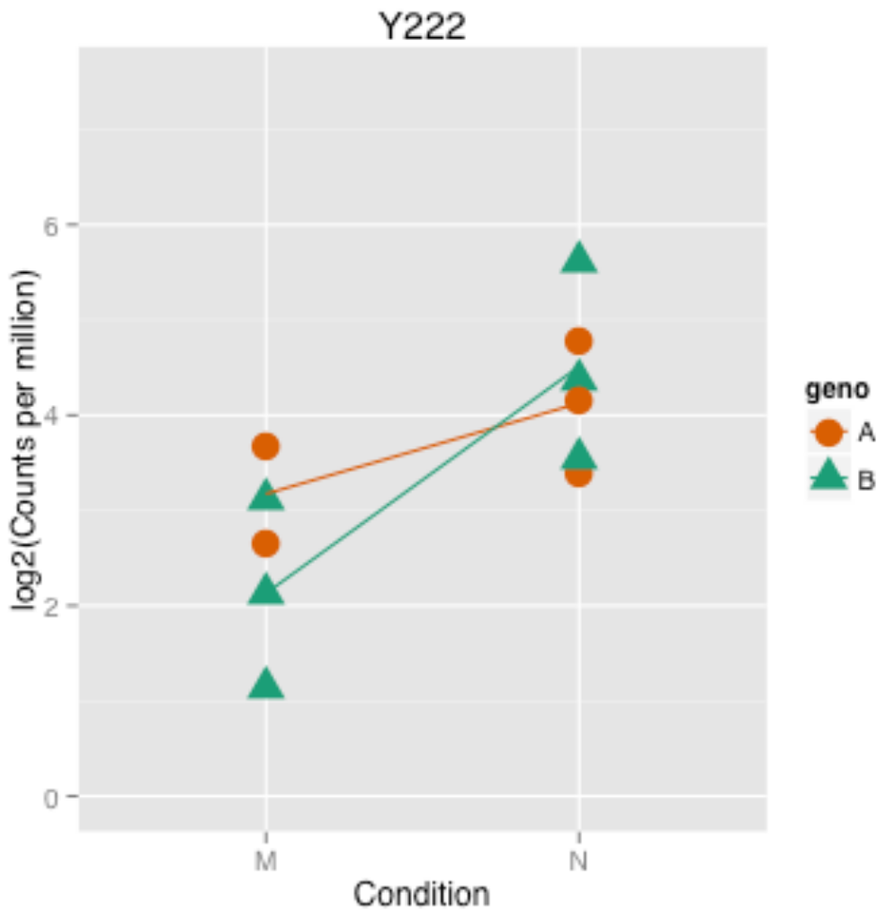
The Good (✔), Maybe (?) & Ugly (✘) ordered list of genes

Do you agree?

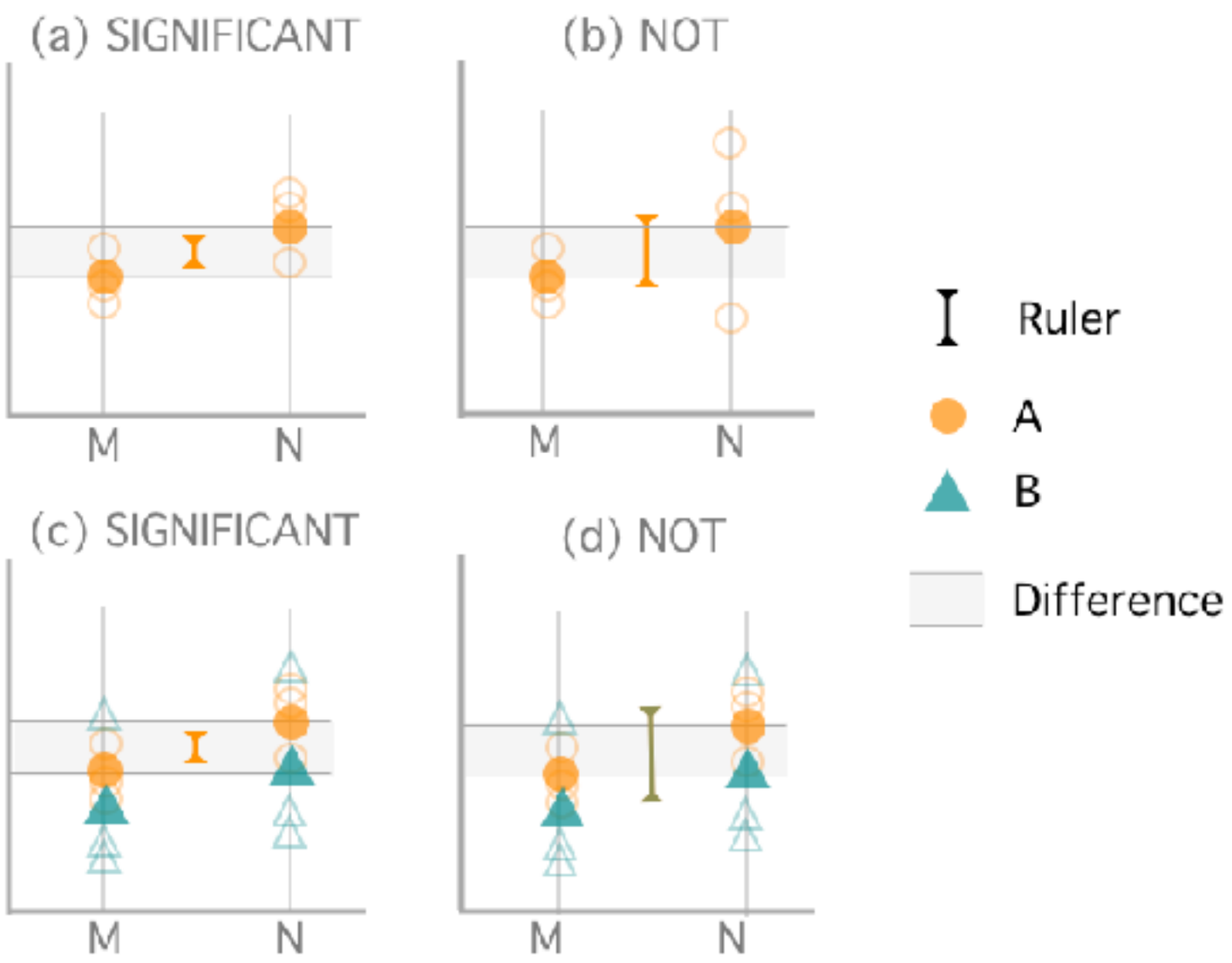
Why?

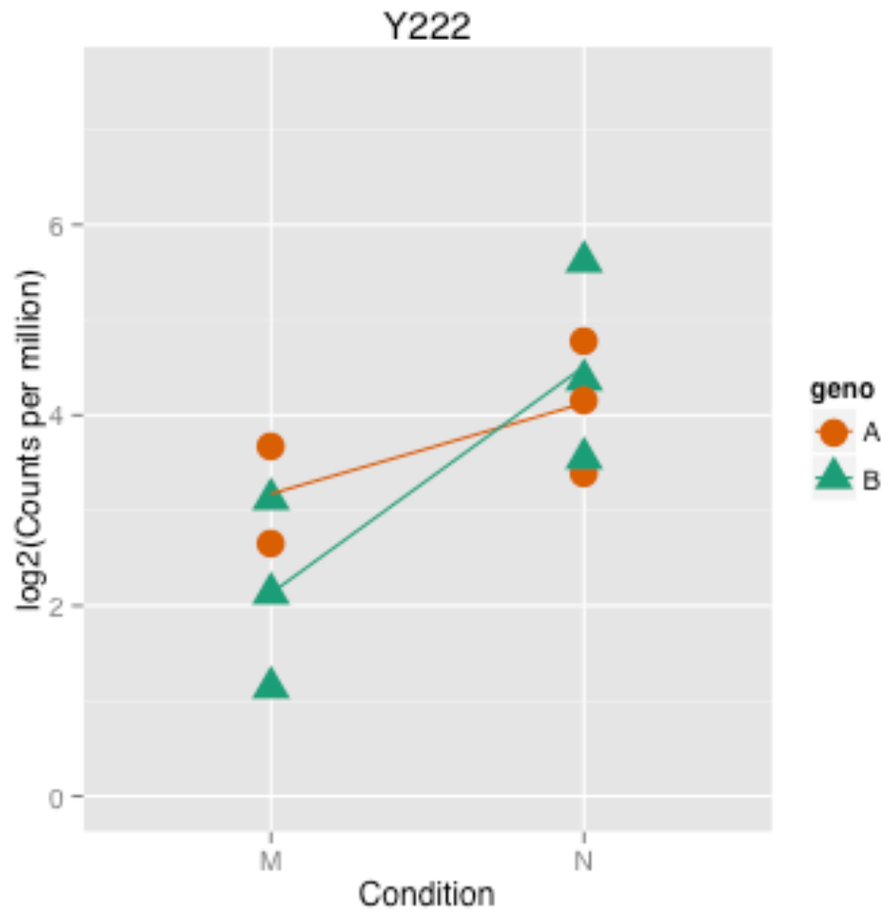
Dispersion



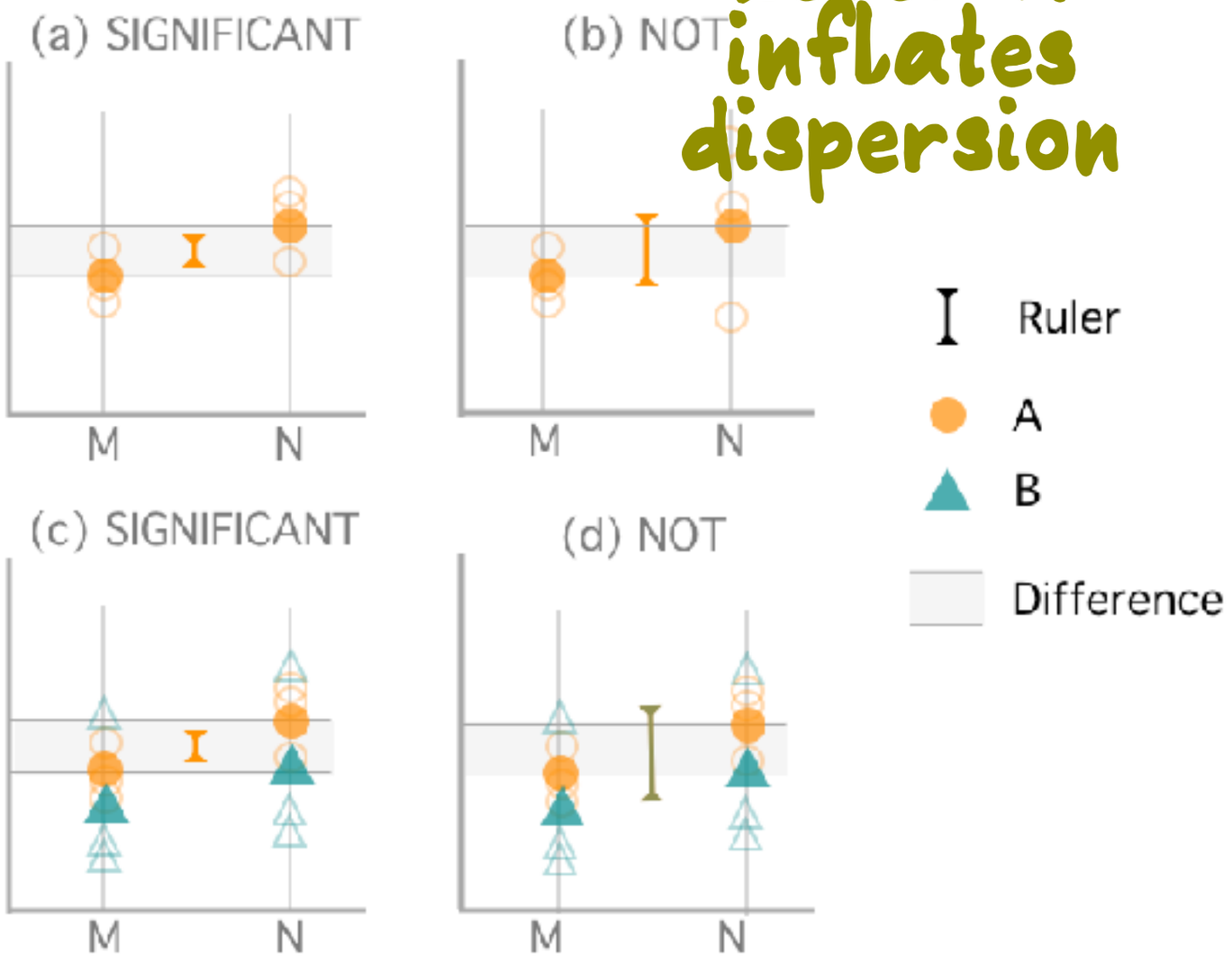


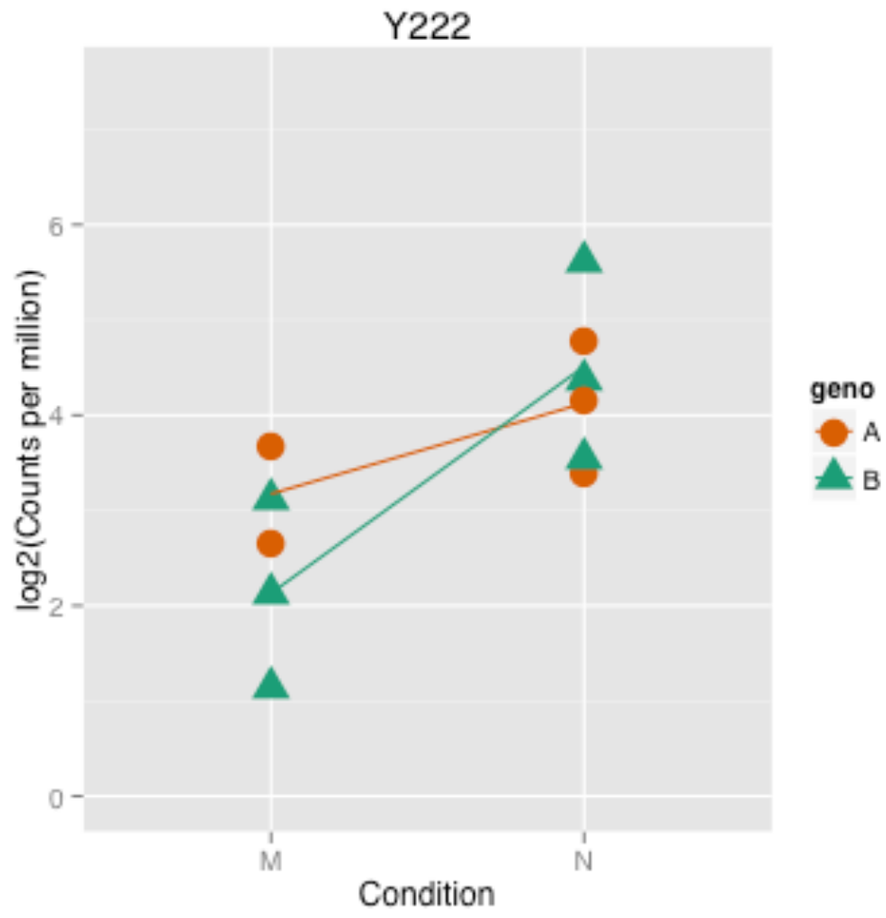
Why?


(a) SIGNIFICANT
(b) NOT
(c) SIGNIFICANT
(d) NOT
M
N
Ruler
A
B
Difference

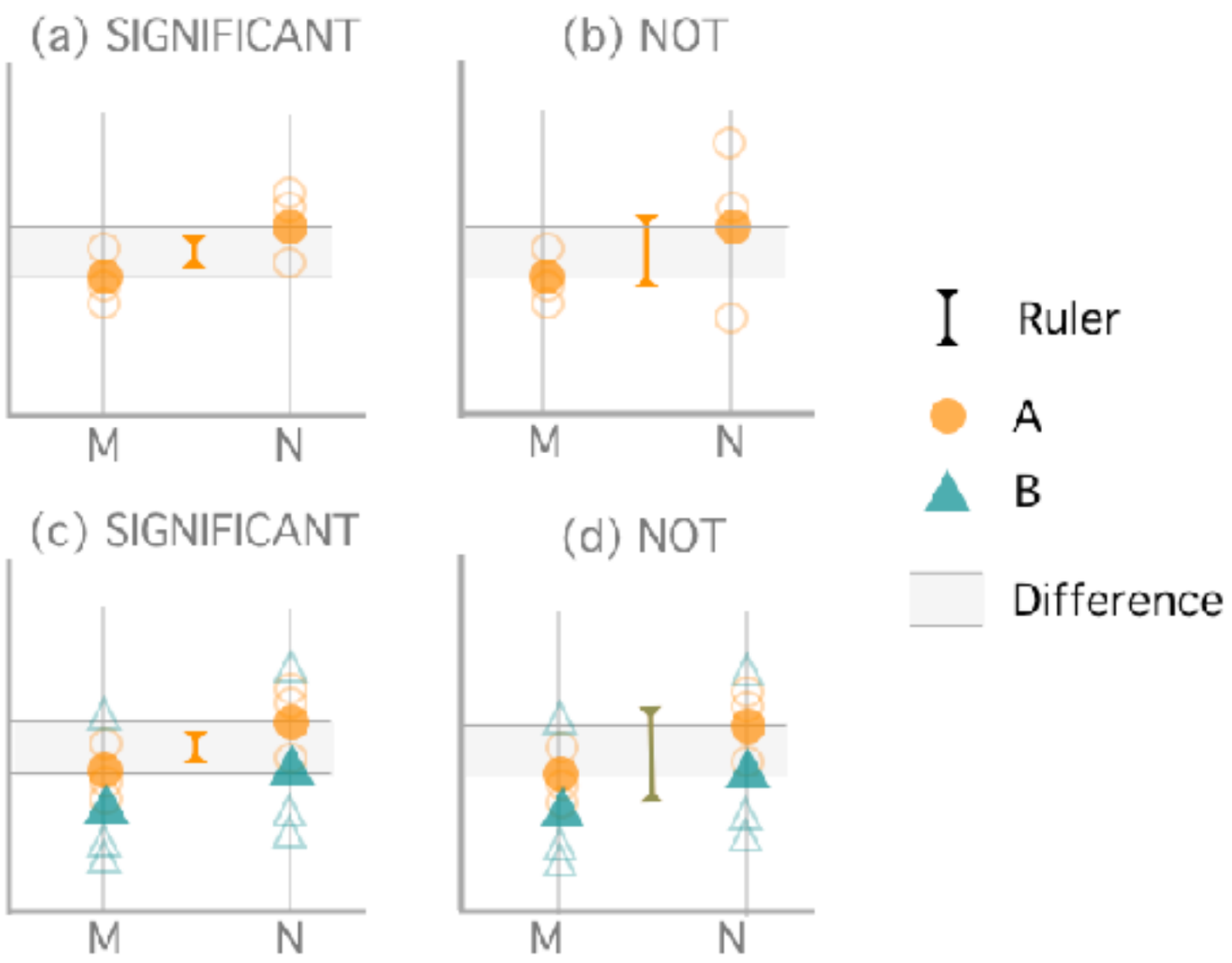

Y222
log2(Counts per million)
6
4
2
0
M
N
Condition
geno
A
B

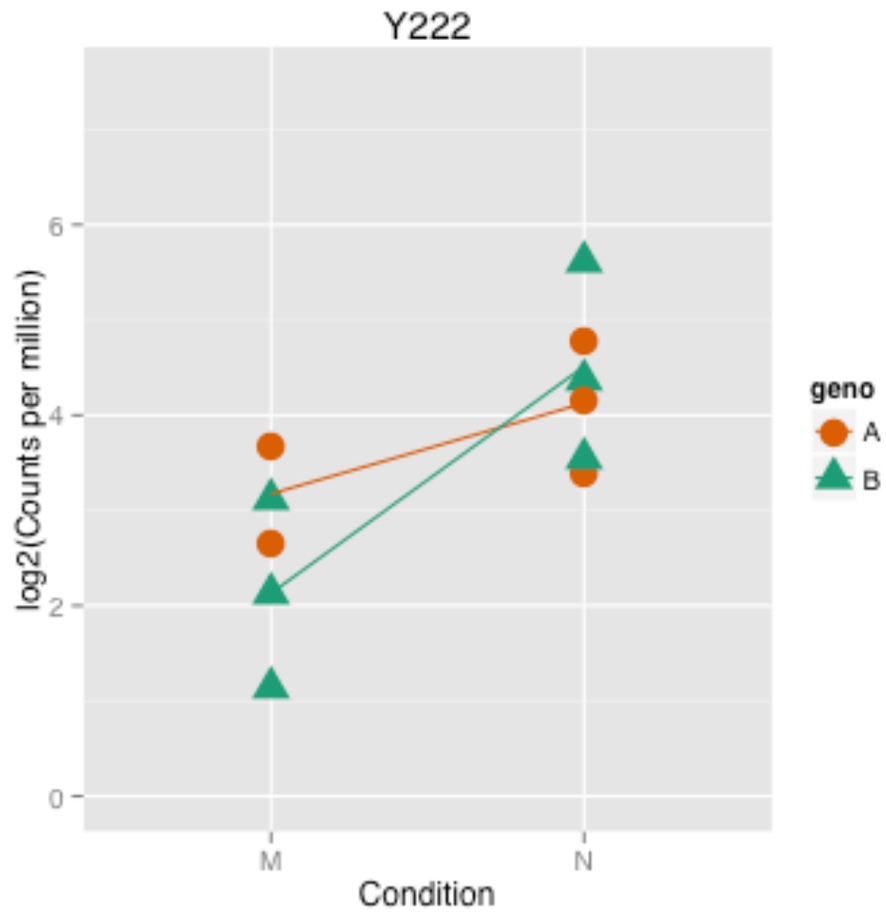
Why?


(a) SIGNIFICANT
(b) NOT
(c) SIGNIFICANT
(d) NOT
M
N
Level N inflates dispersion
Ruler
A
B
Difference

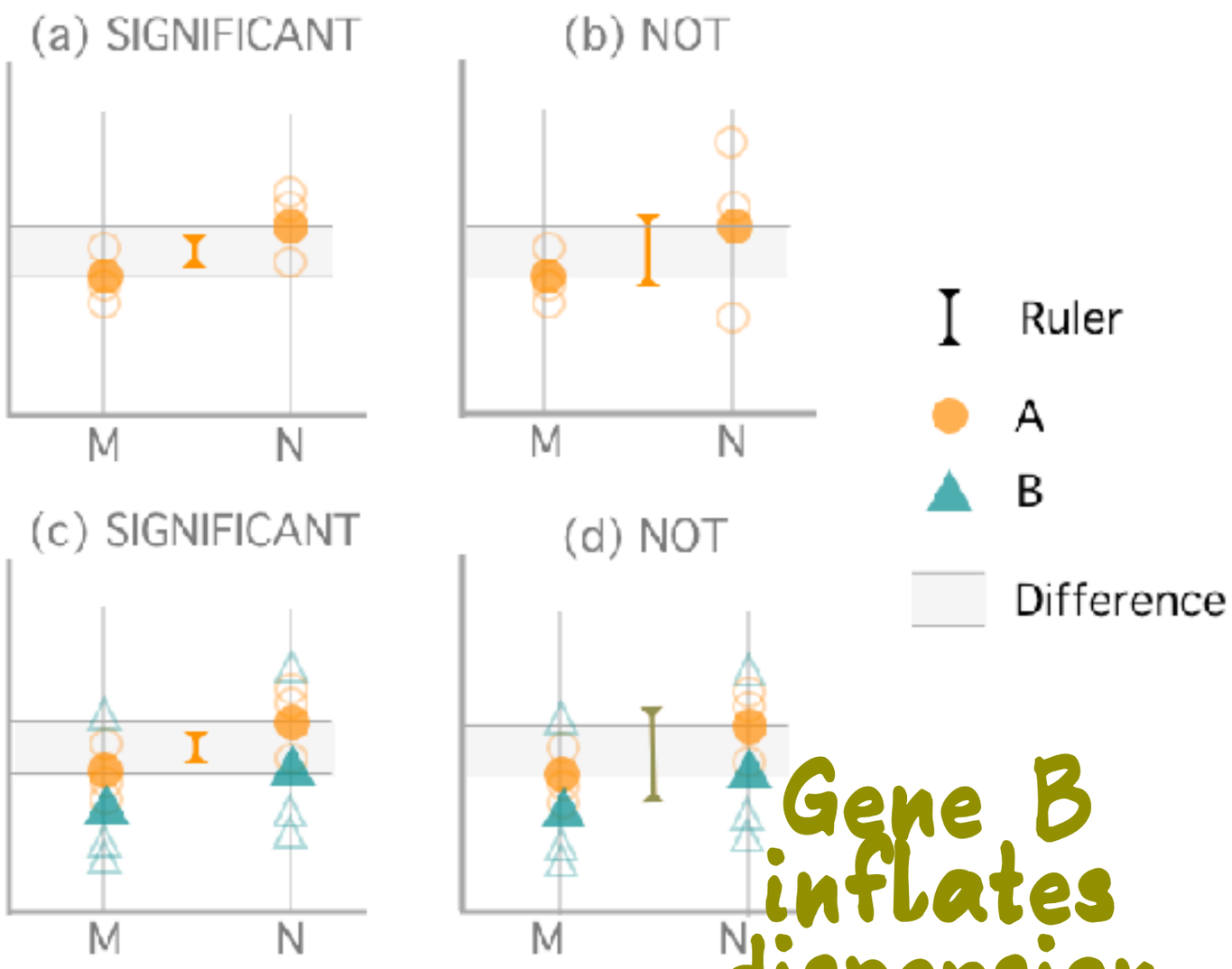

Y222
log2(Counts per million)
6
4
2
0
M
N
Condition
geno
A
B

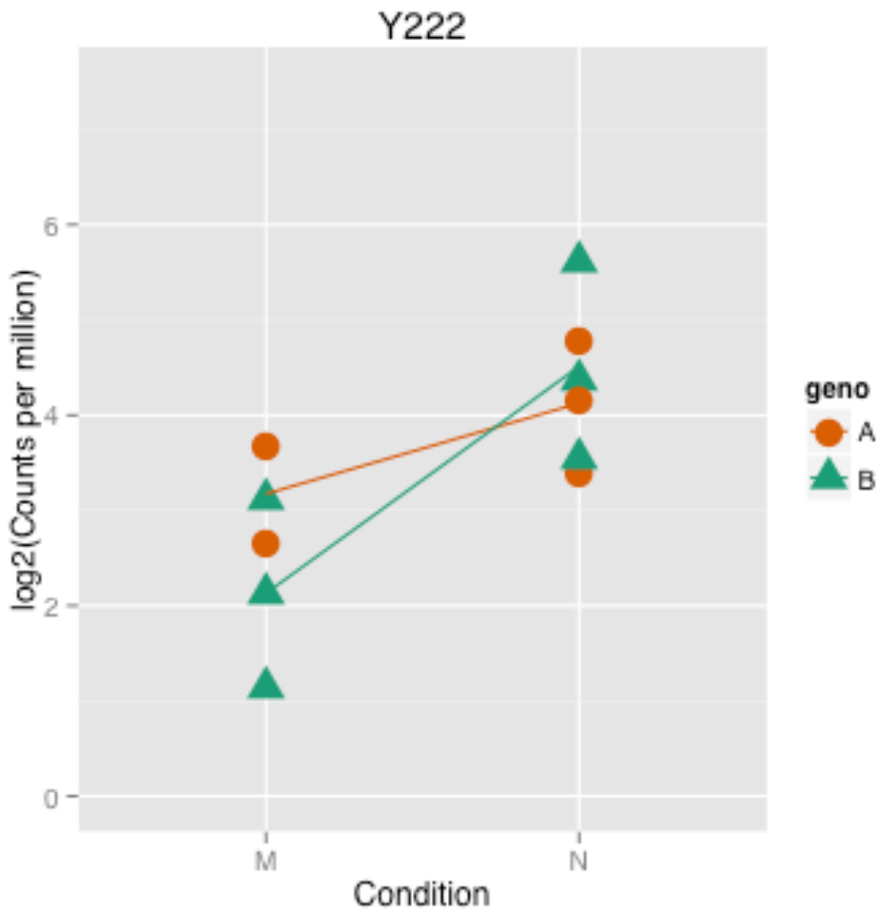
Why?


(a) SIGNIFICANT
(b) NOT
(c) SIGNIFICANT
(d) NOT
M
N
Ruler
A
B
Difference


Y222
log2(Counts per million)
6
4
2
0
M
N
Condition
geno
A
B

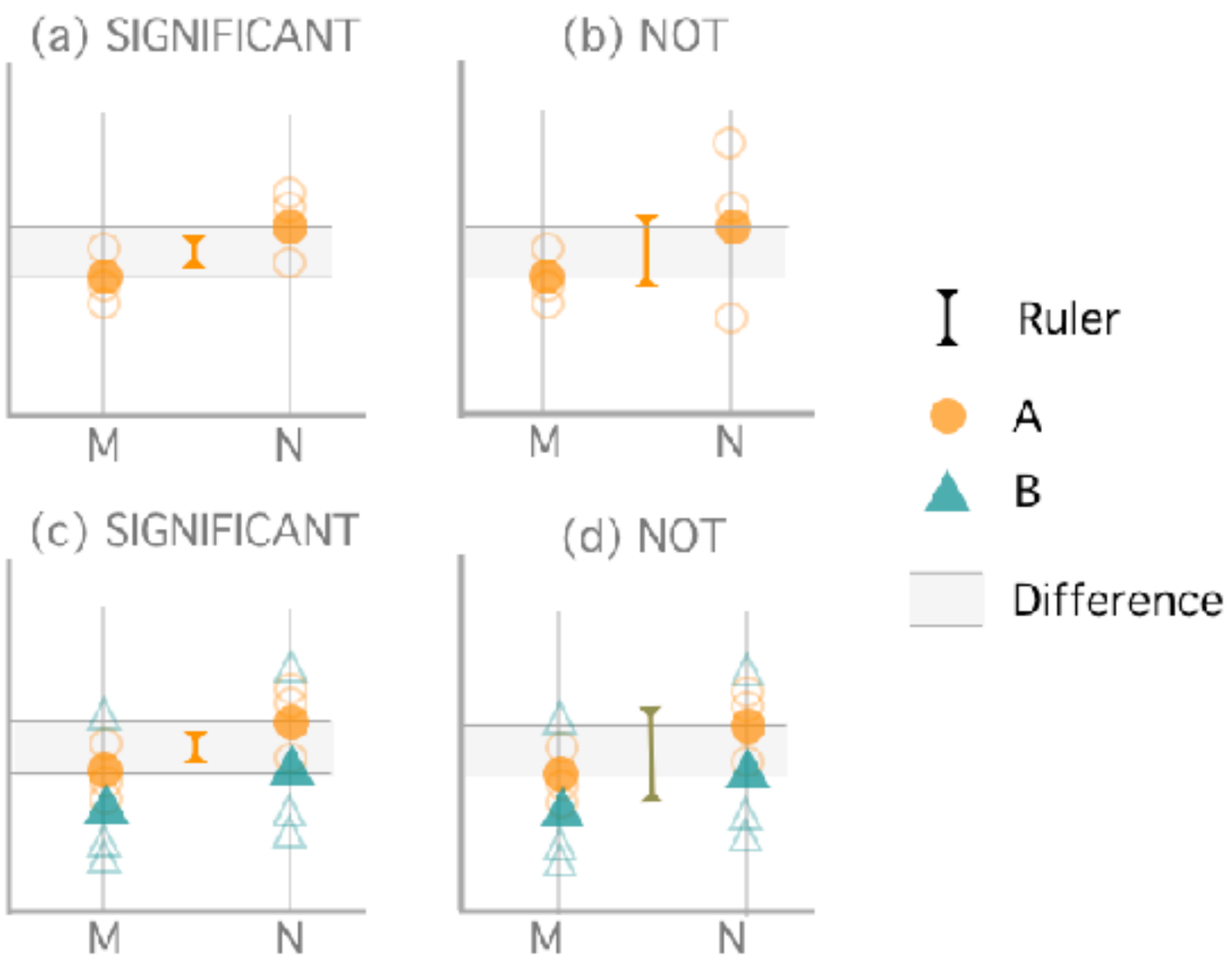
Why?

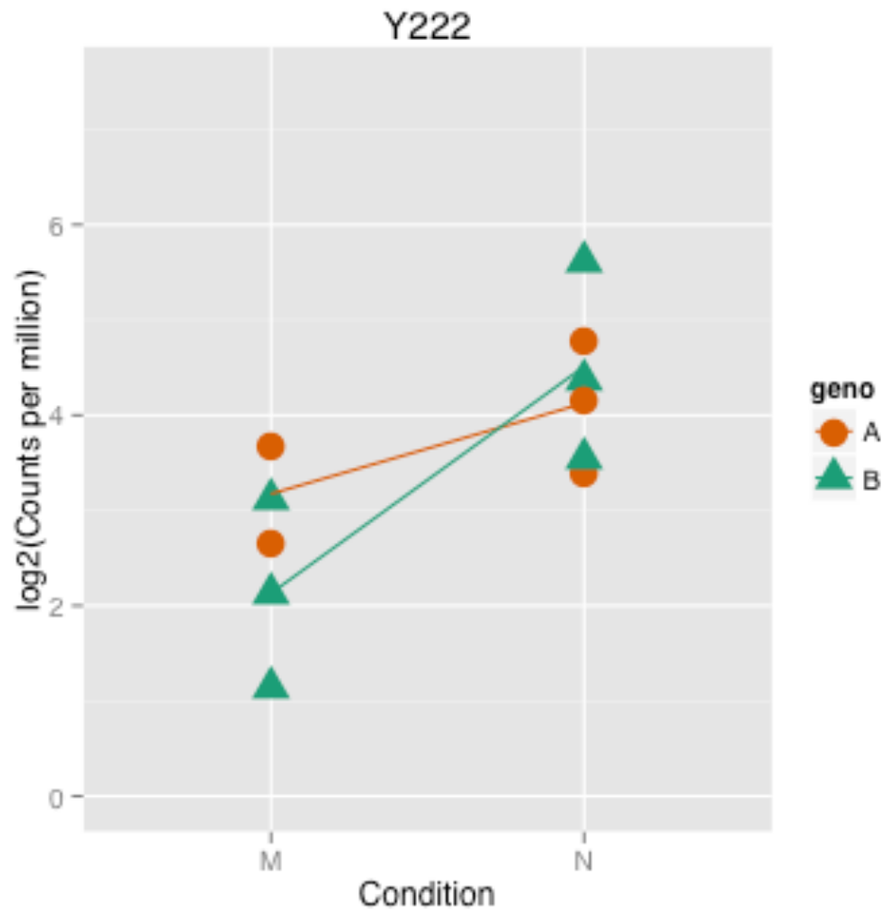




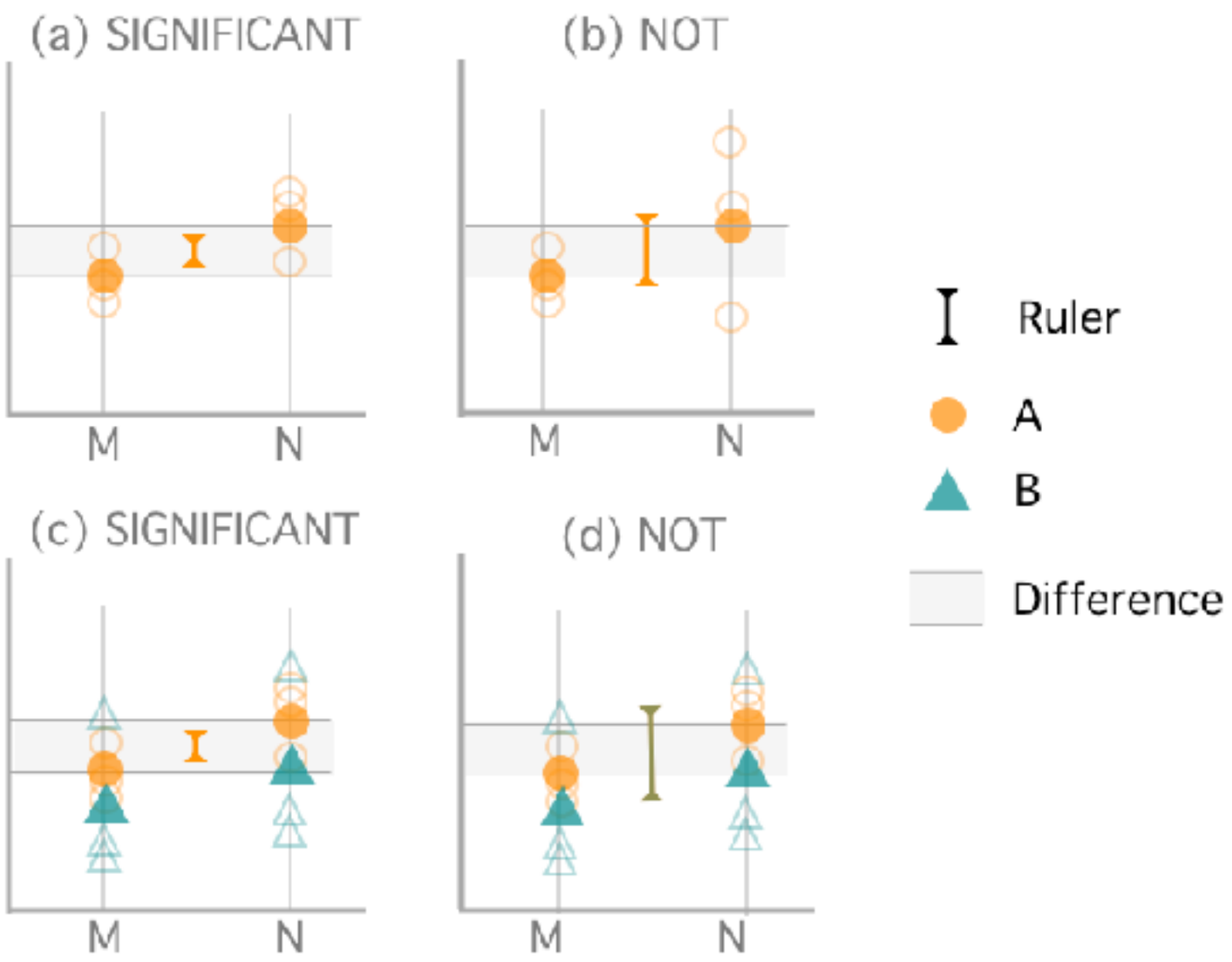
Gene B inflates dispersion

Why?

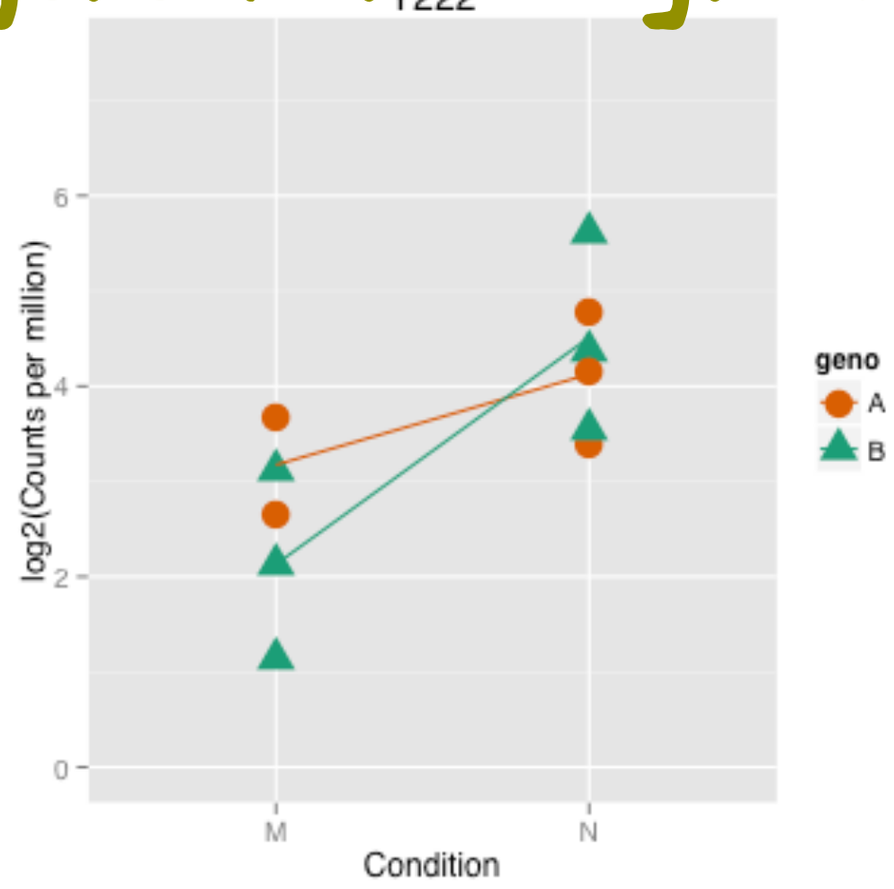

(a) SIGNIFICANT
(b) NOT
(c) SIGNIFICANT
(d) NOT
M
N
Ruler
A
B
Difference


Y222
log2(Counts per million)
6
4
2
0
M
N
Condition
geno
A
B

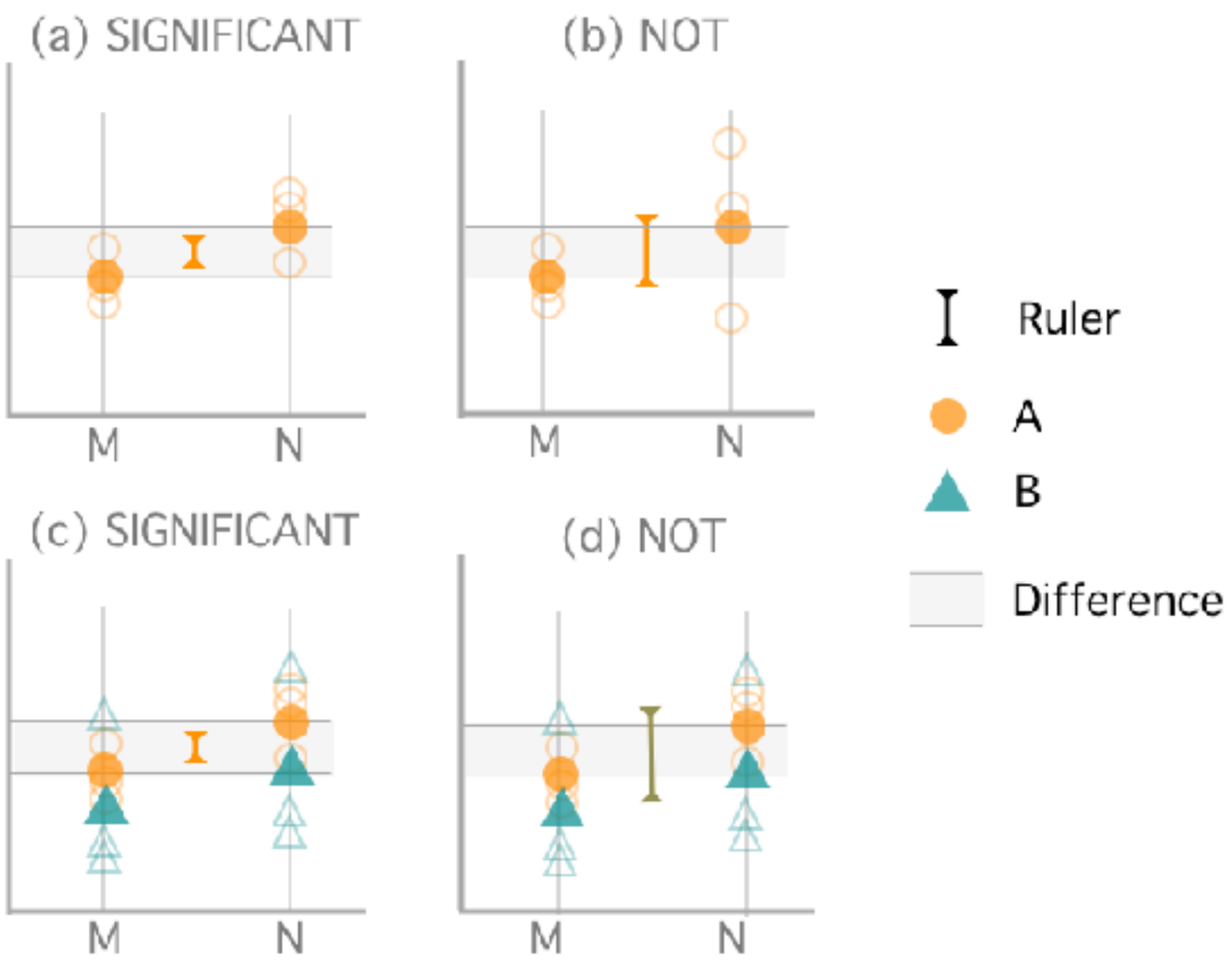
Why?

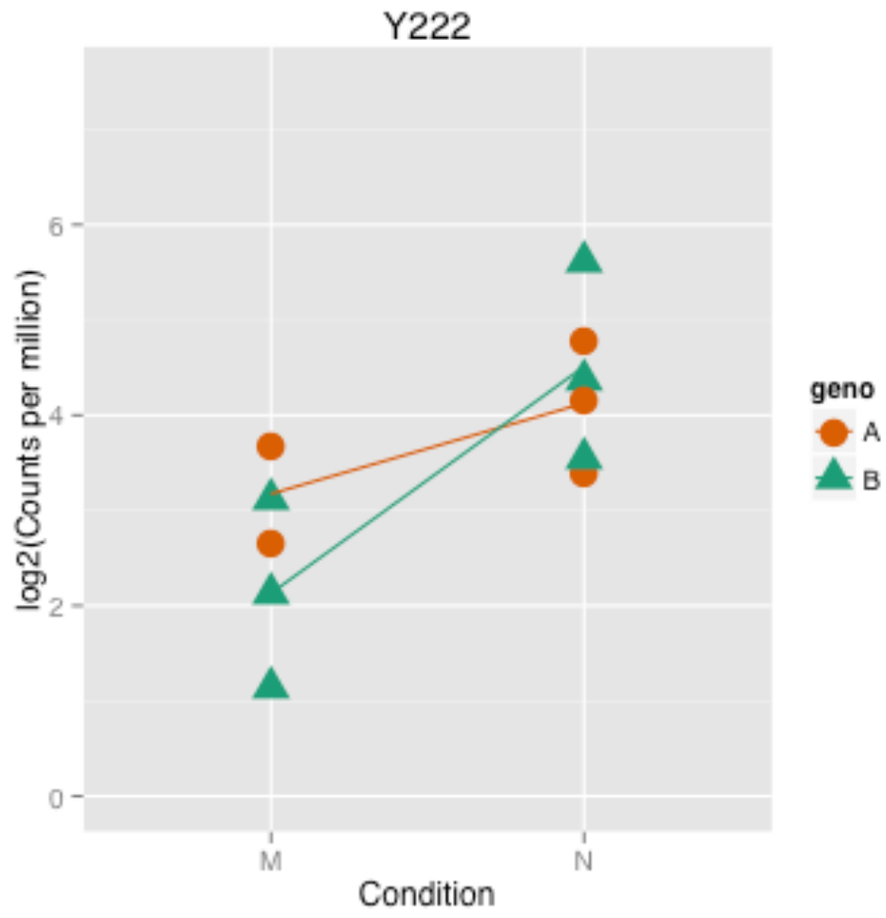


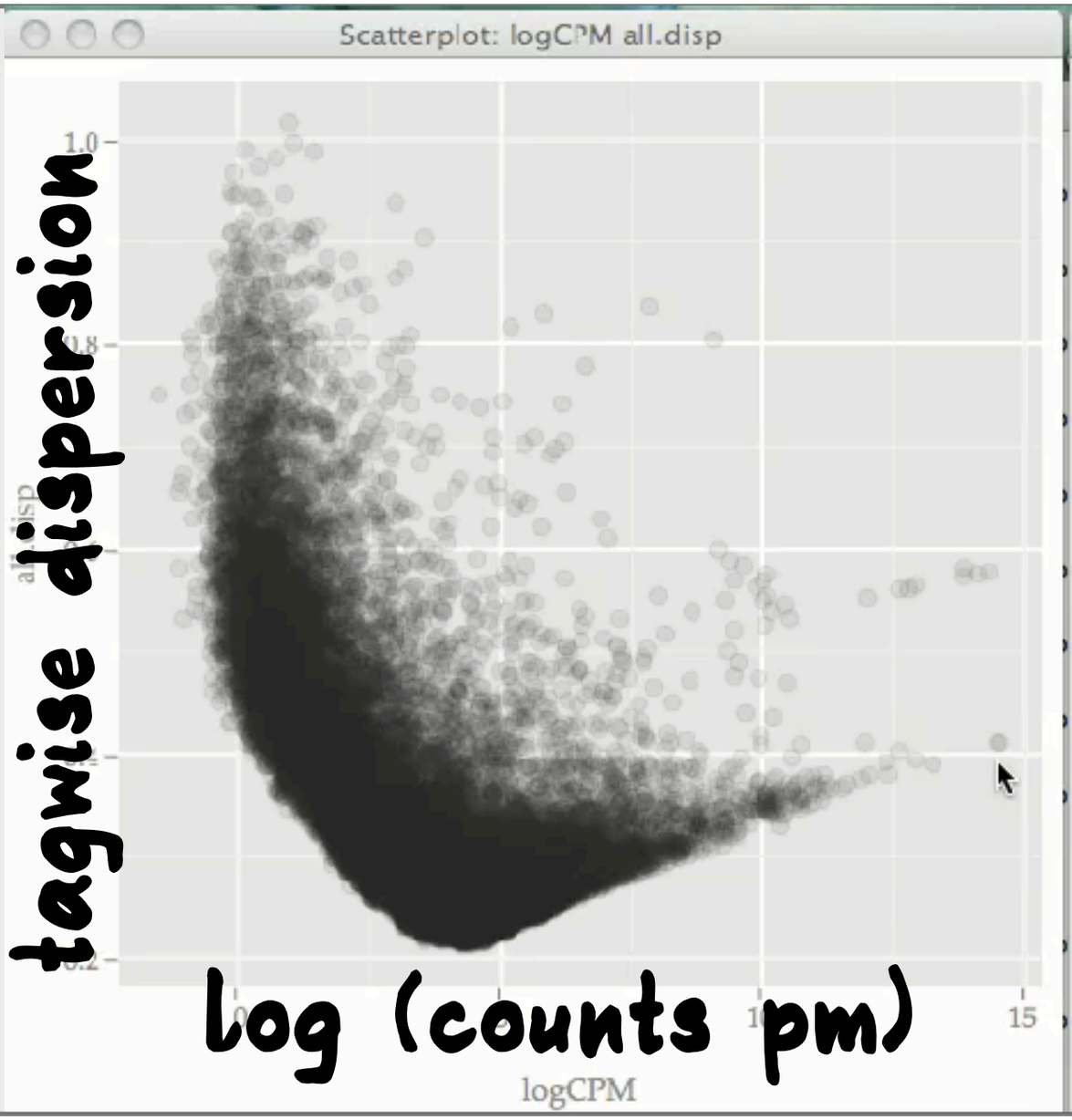
In reality, gene B here inflates dispersion, making gene A not signif.

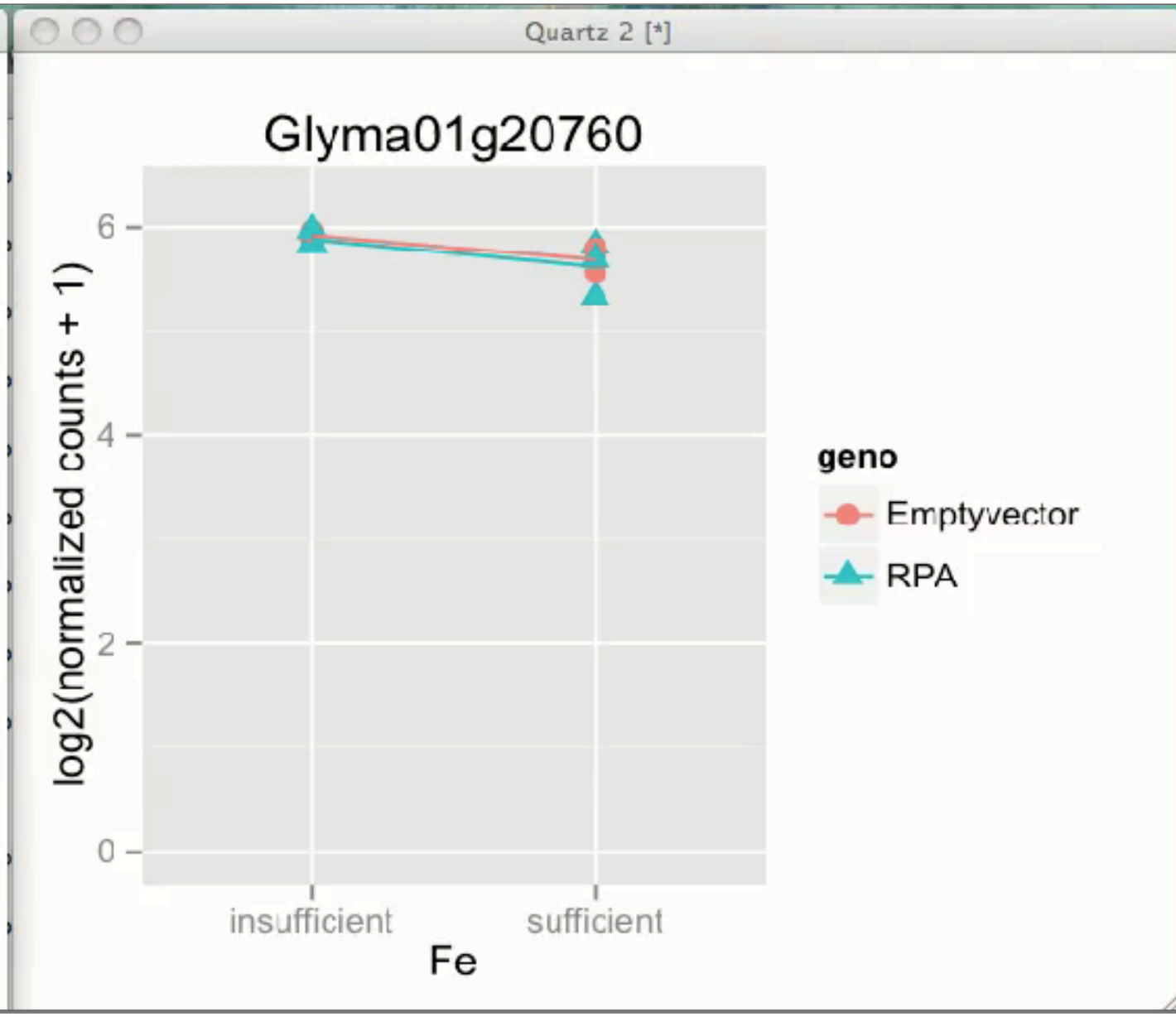


Why?


(a) SIGNIFICANT
(b) NOT
(c) SIGNIFICANT
(d) NOT
M
N
Ruler
A
B
Difference

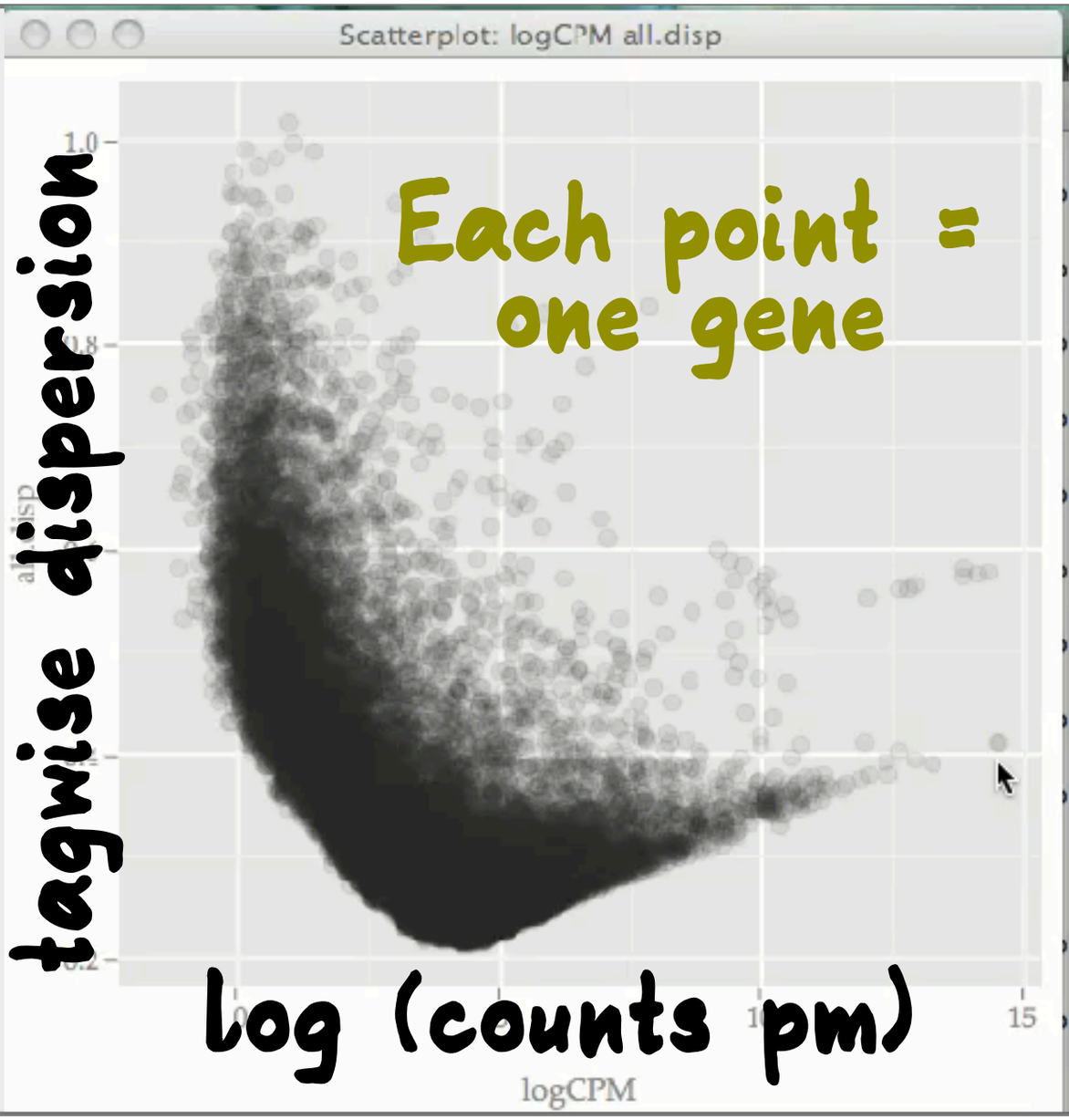

Y222
log2(Counts per million)
6
4
2
0
M
N
Condition
geno
A
B

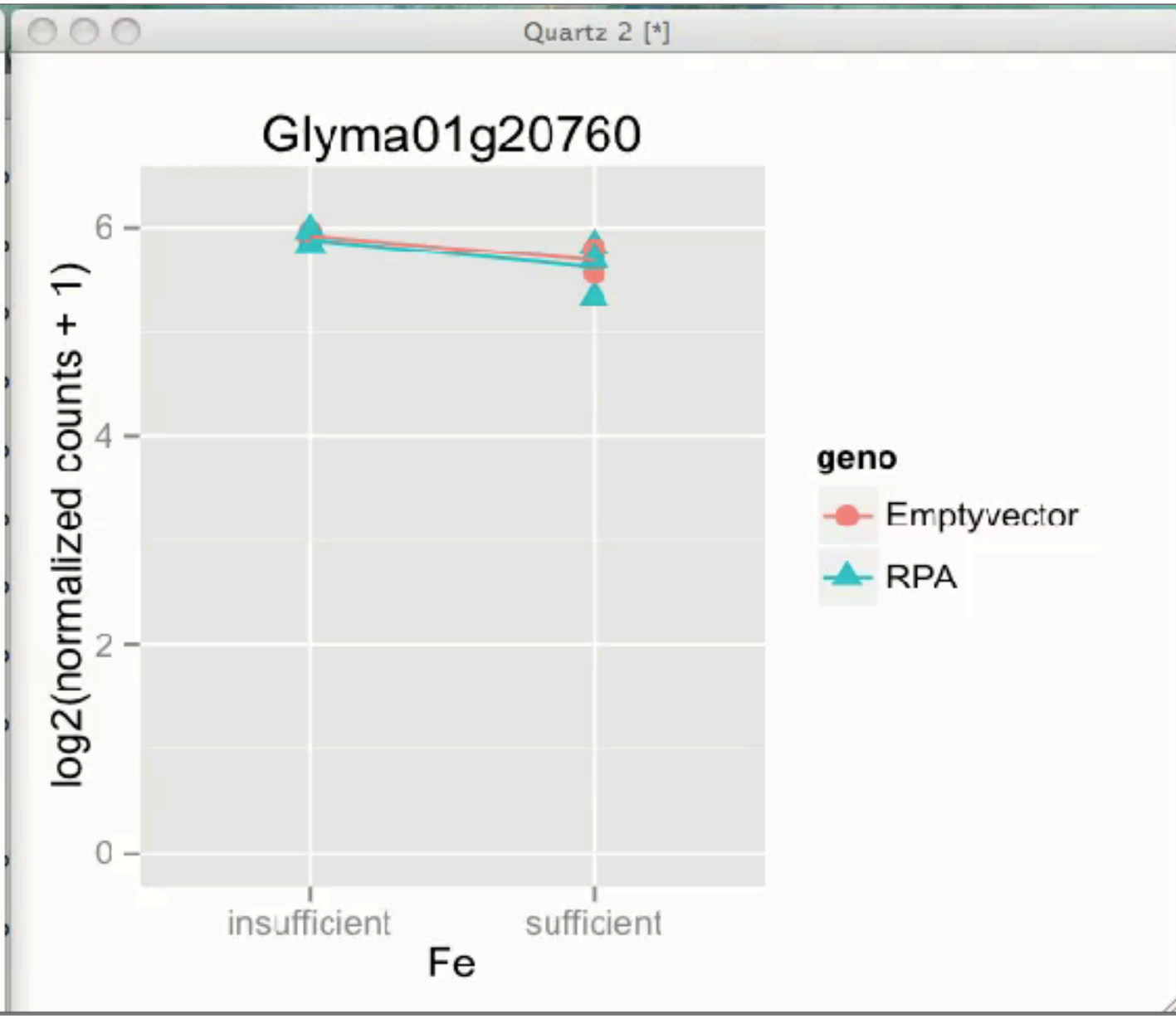




cranvas

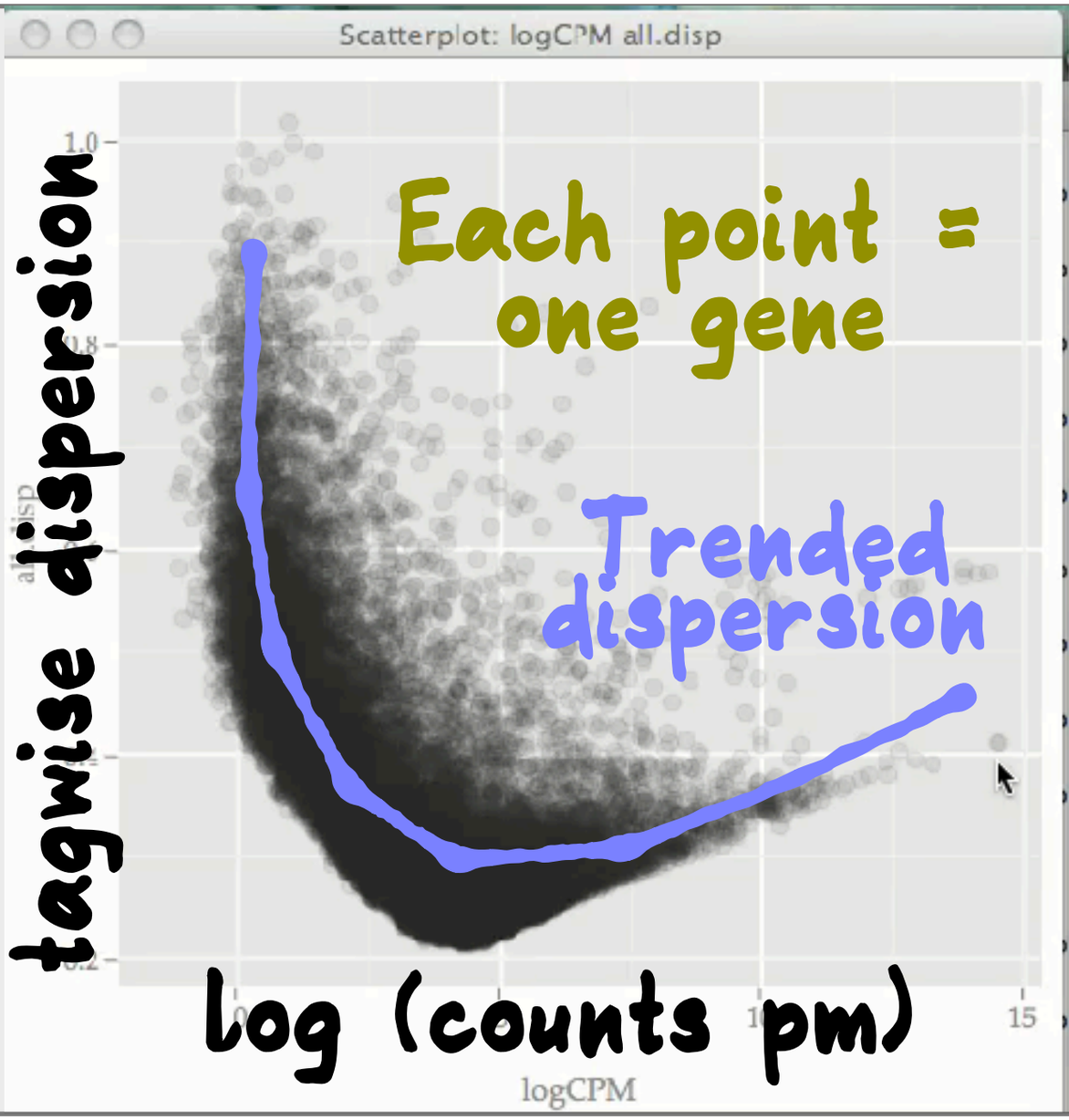
ggplot2

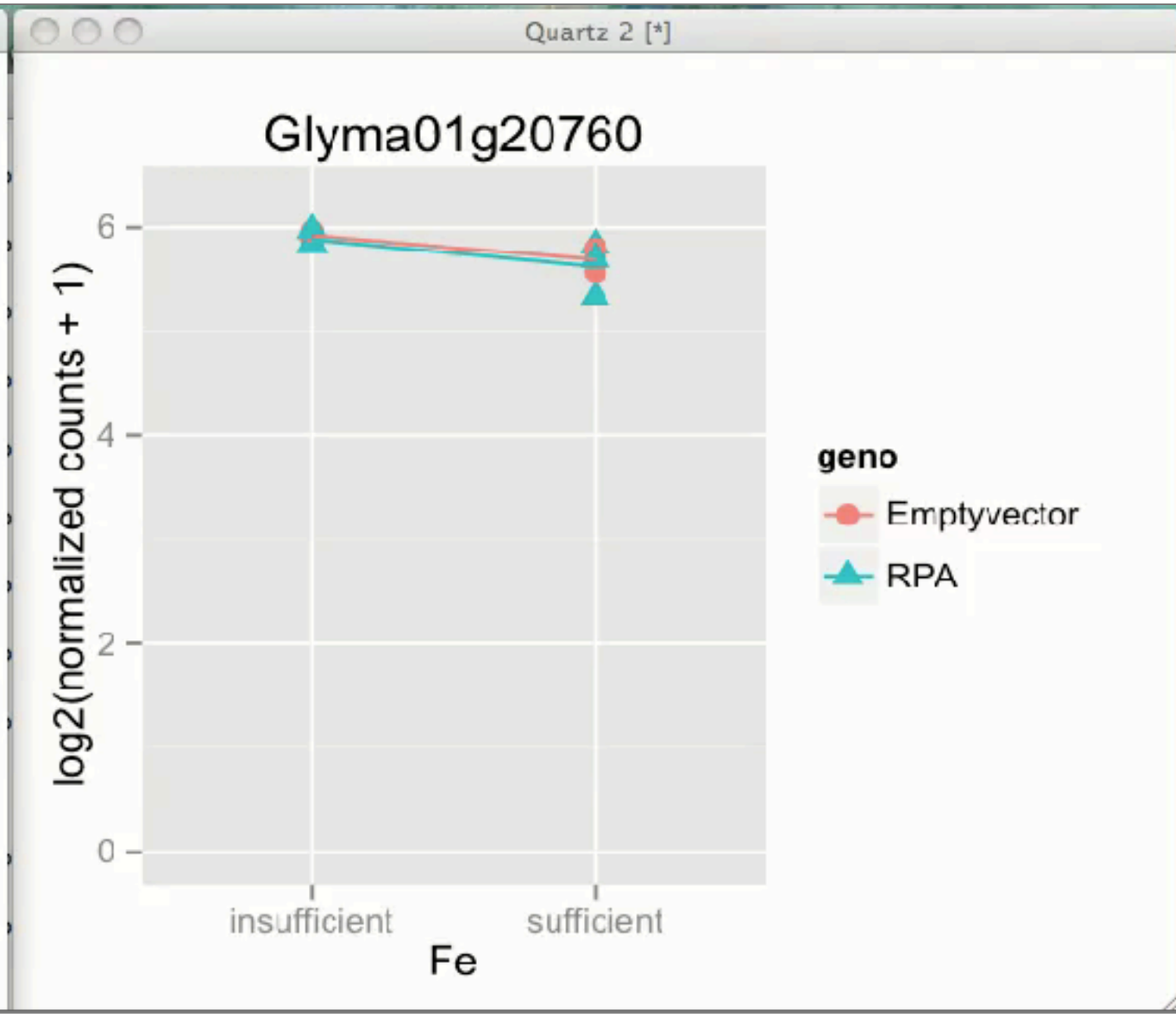




cranvas

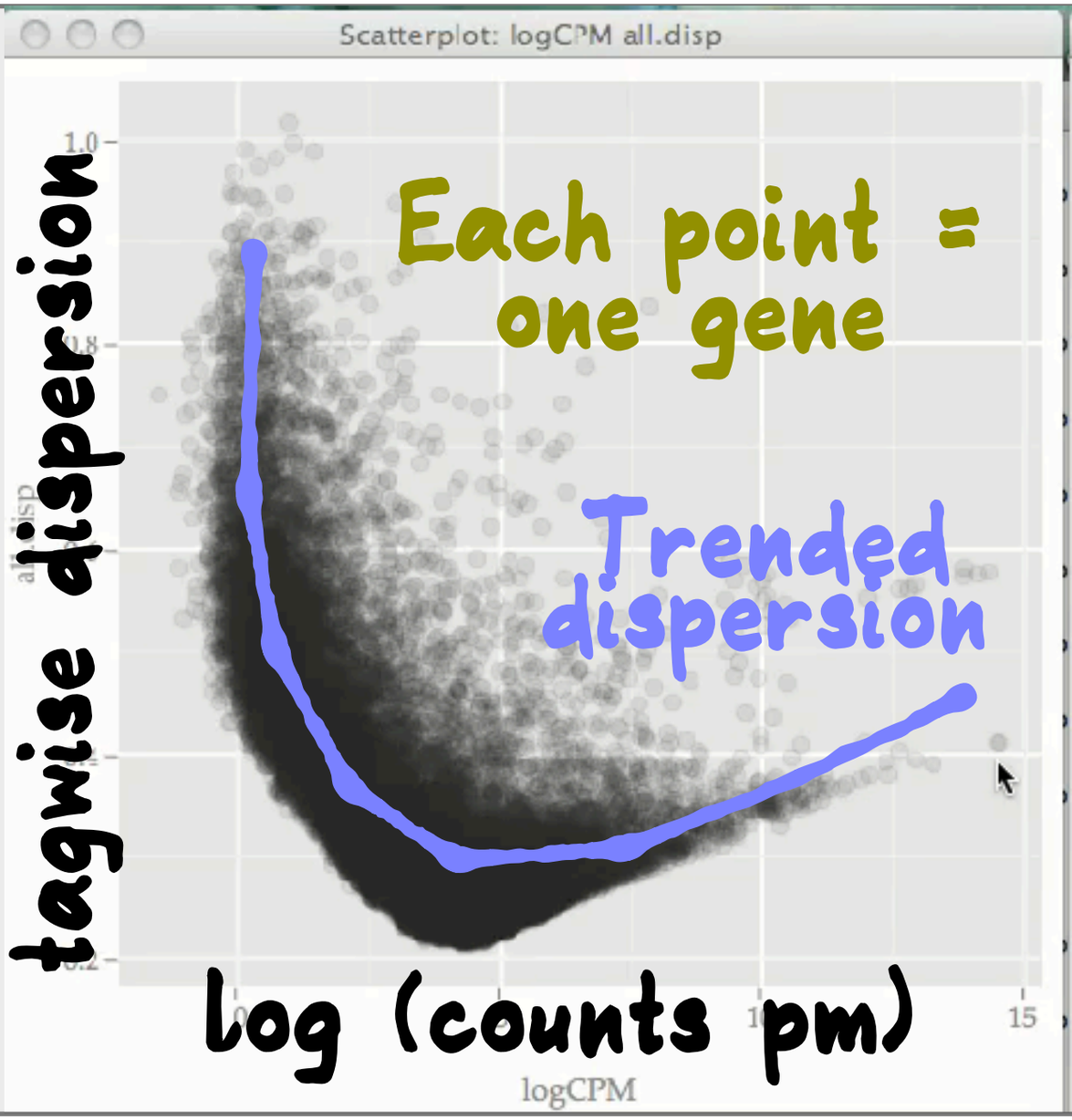
ggplot2





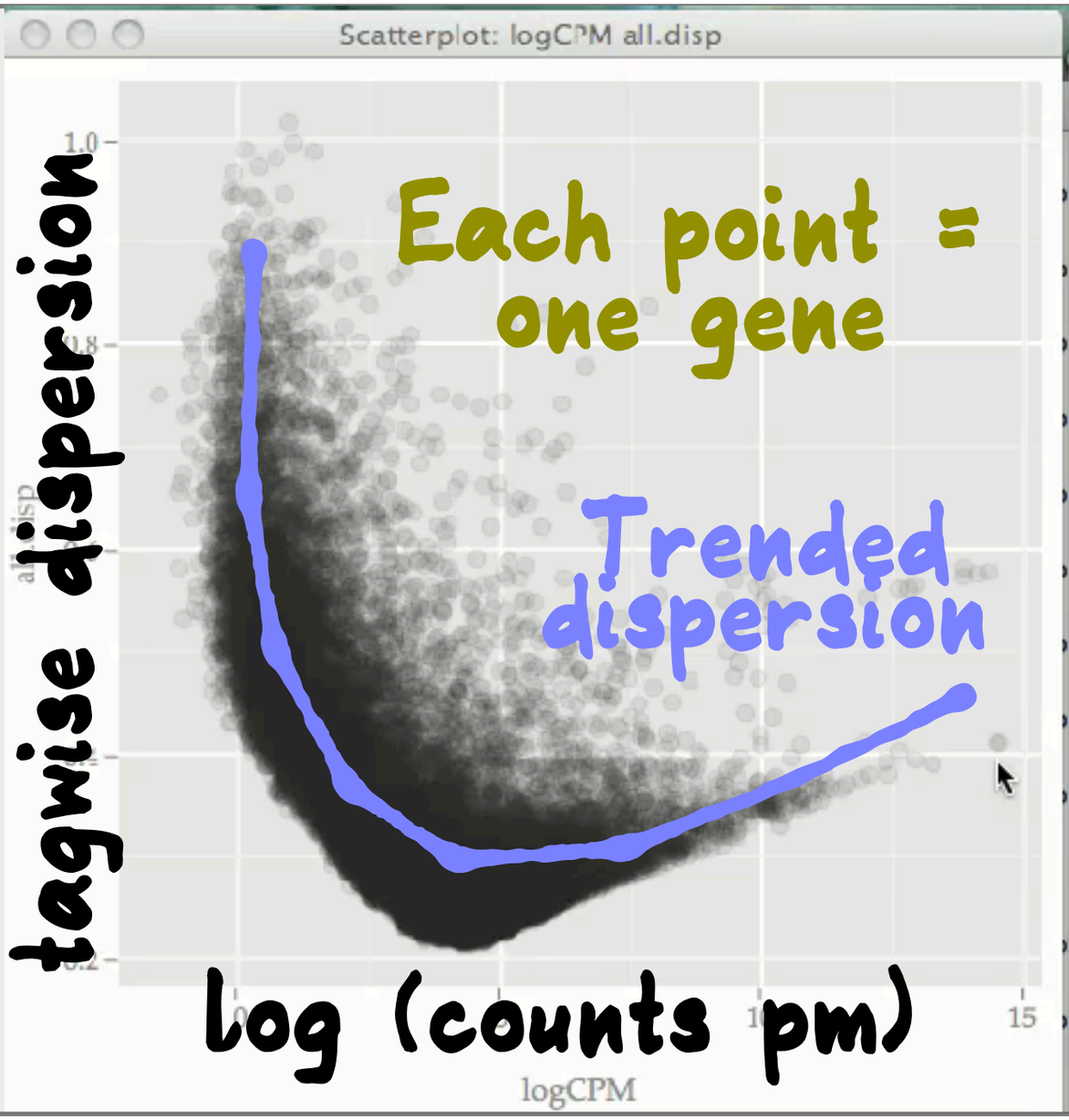
cranvas

ggplot2



cranvas

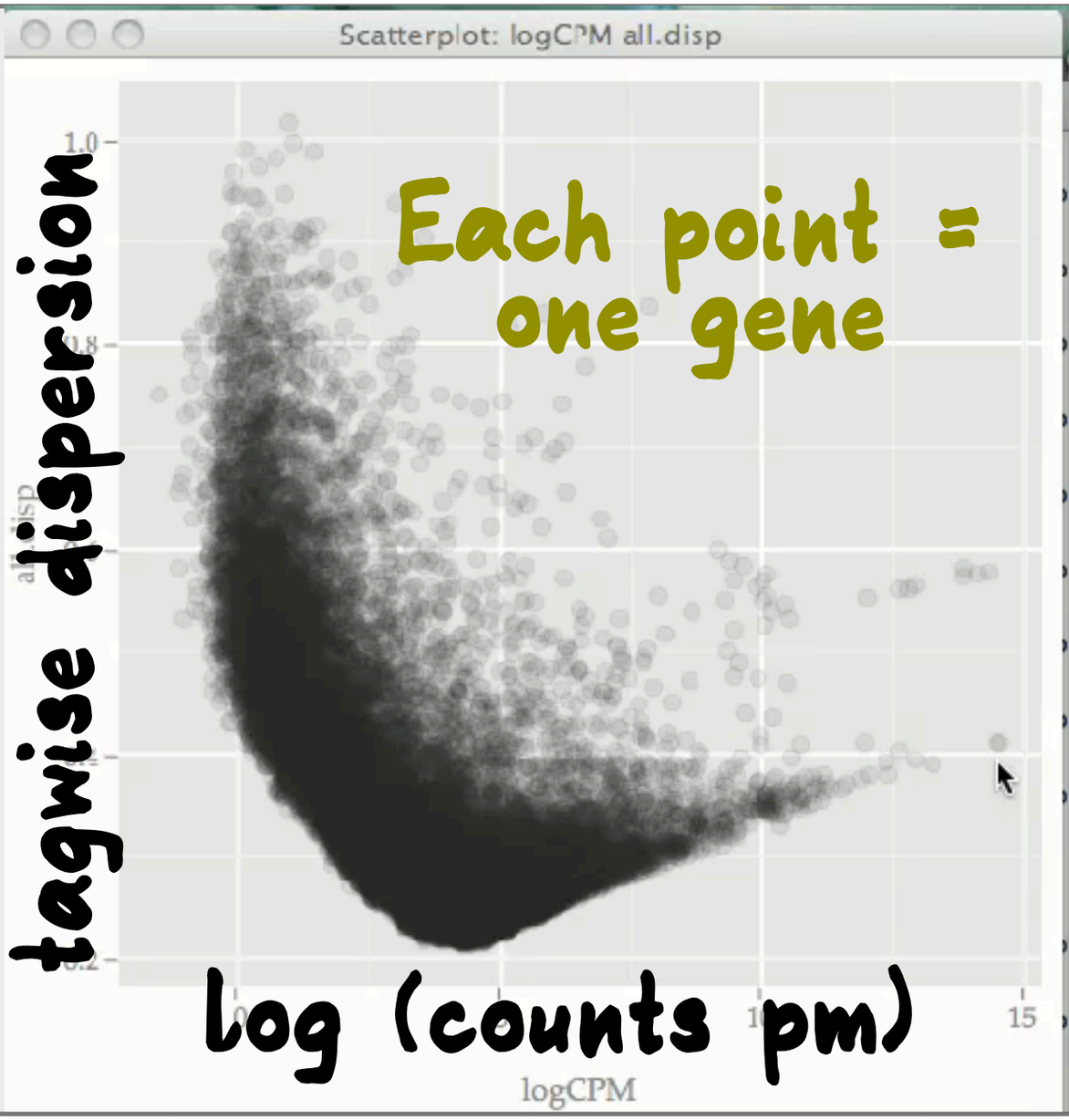
ggplot2



Plots linked, clicking on a point in left plot shows the interaction plot for that gene

cranvas

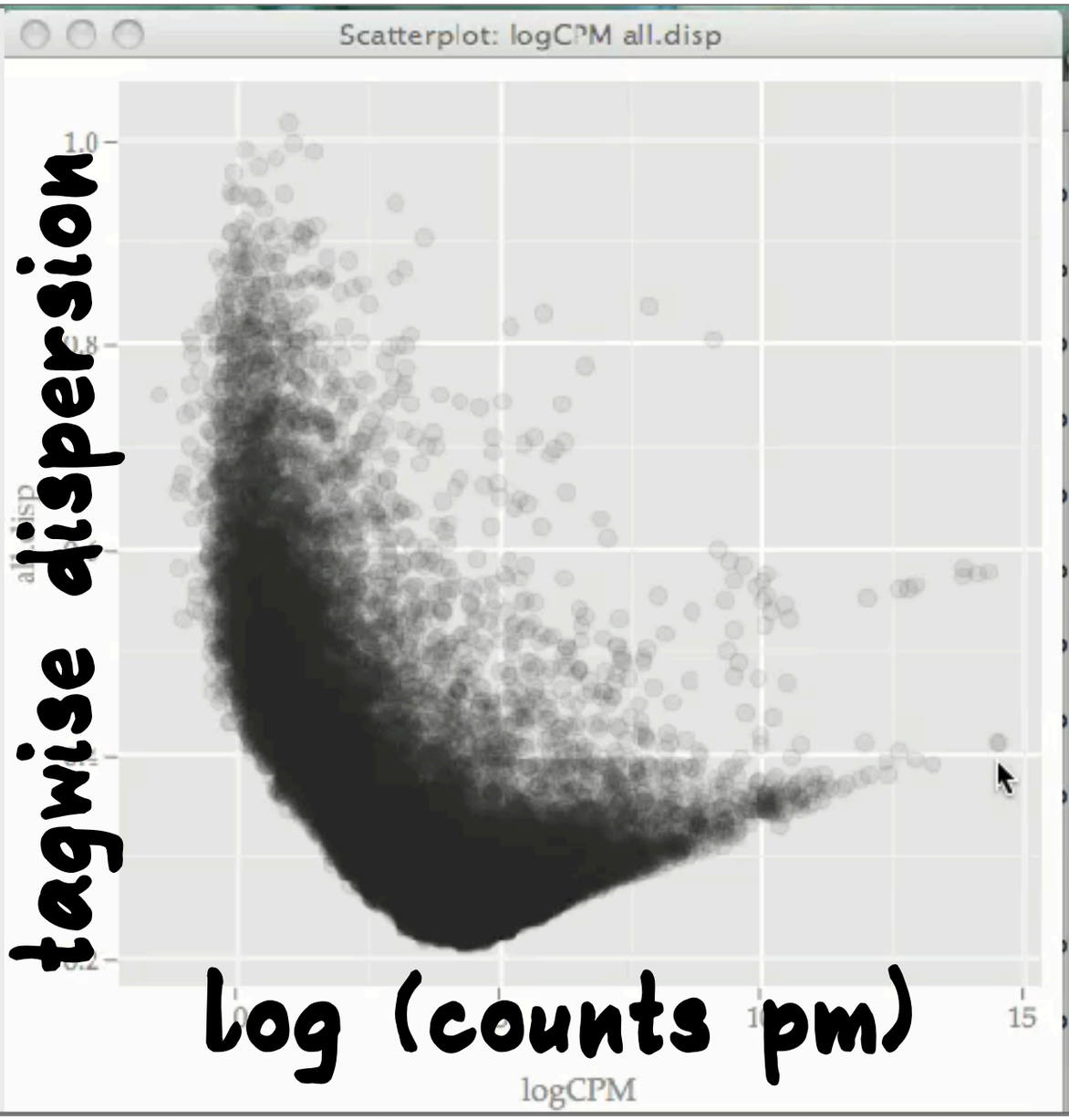
ggplot2



Plots linked, clicking on a point in left plot shows the interaction plot for that gene

cranvas

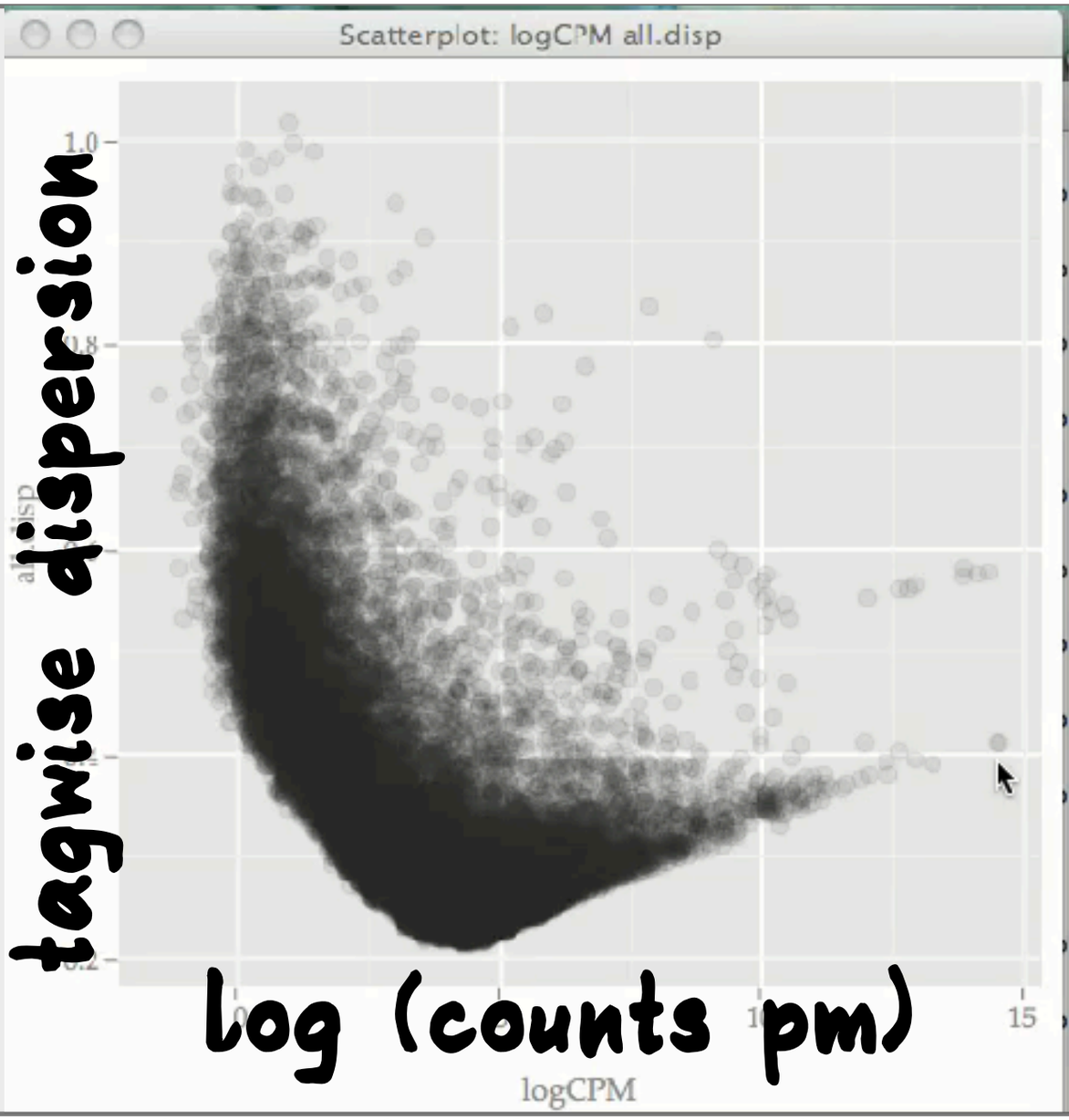
ggplot2



Plots linked, clicking on a point in left plot shows the interaction plot for that gene

cranvas

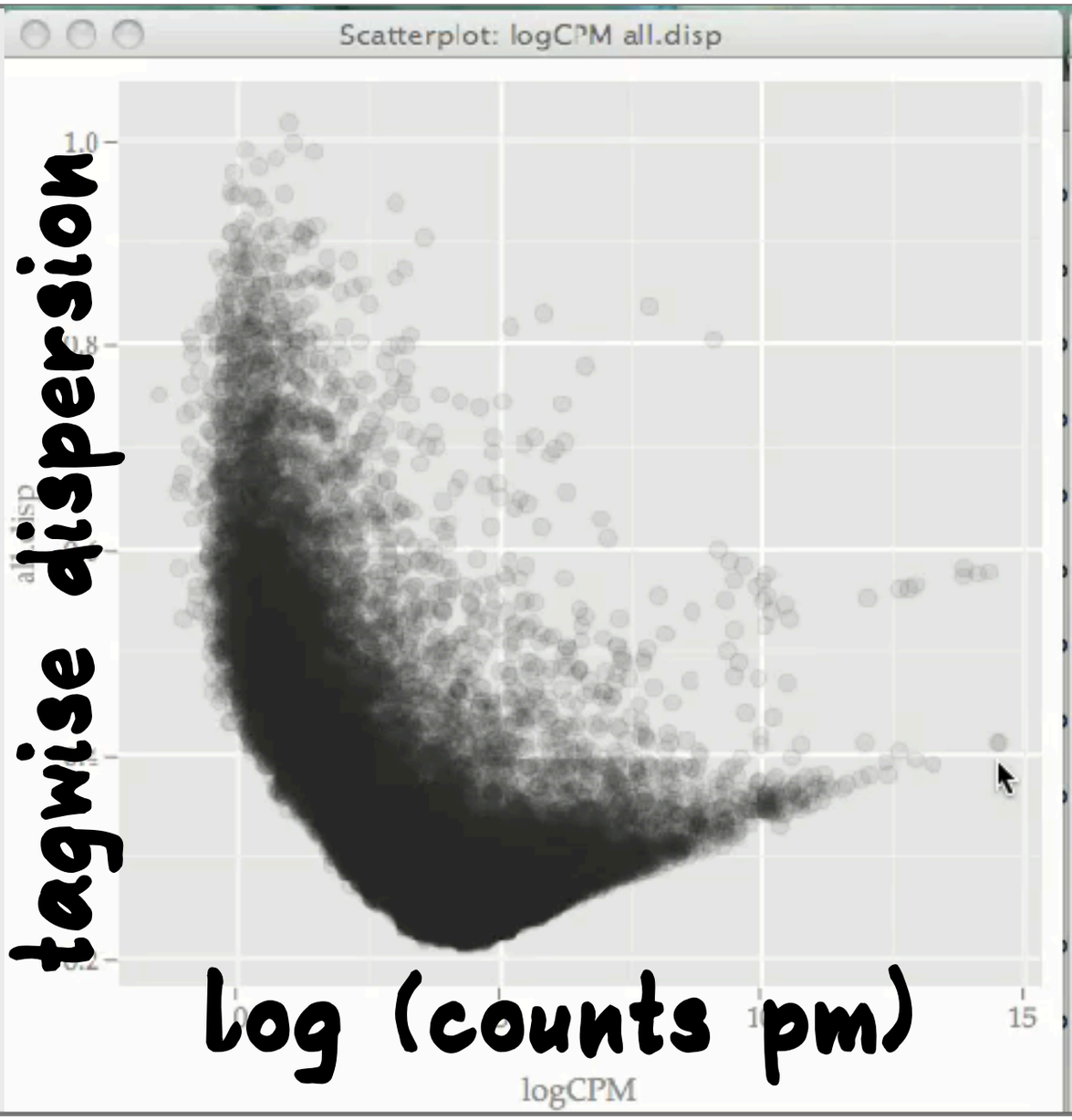
ggplot2

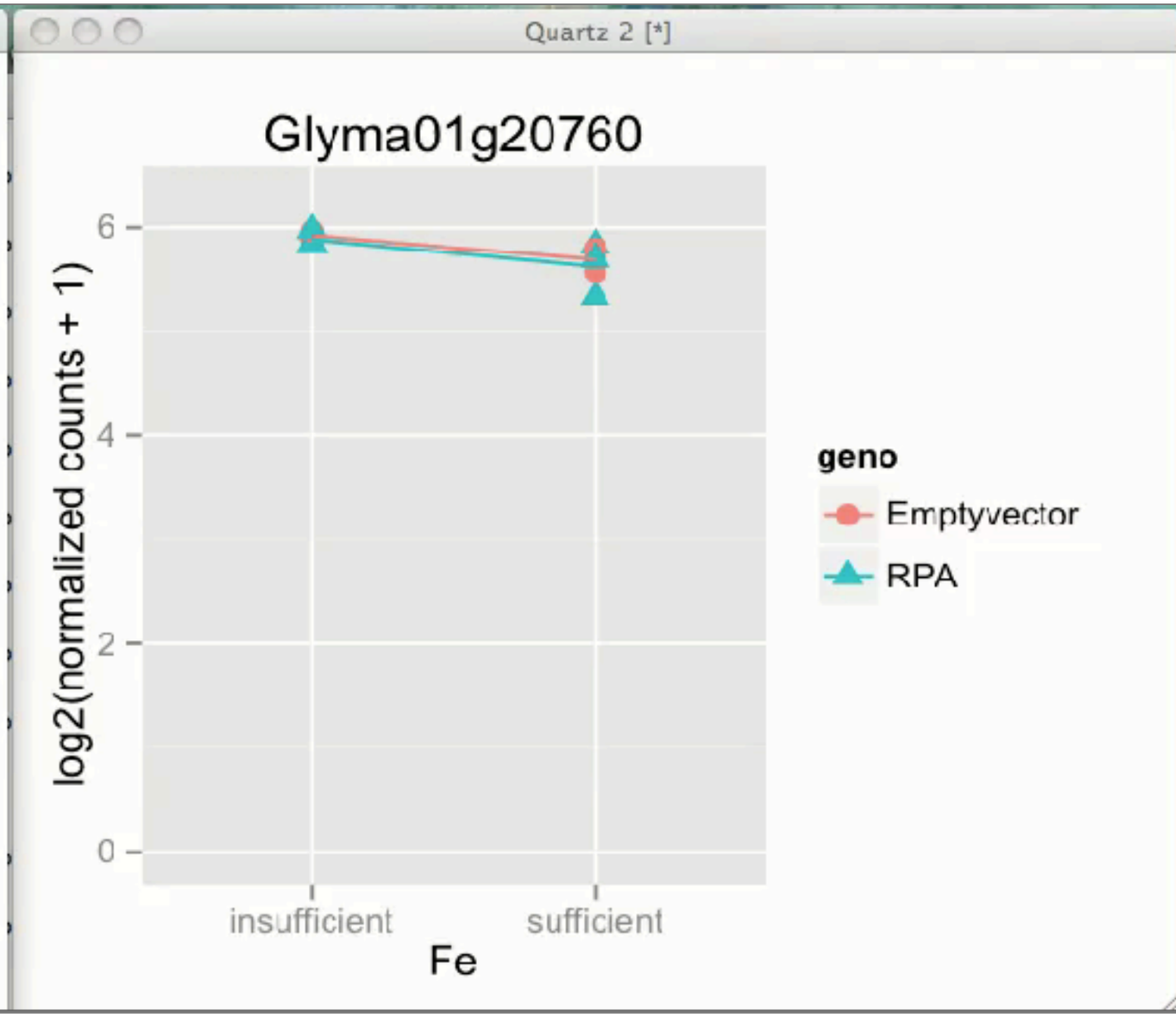


Plots linked, clicking on a point in left plot shows the interaction plot for that gene

cranvas

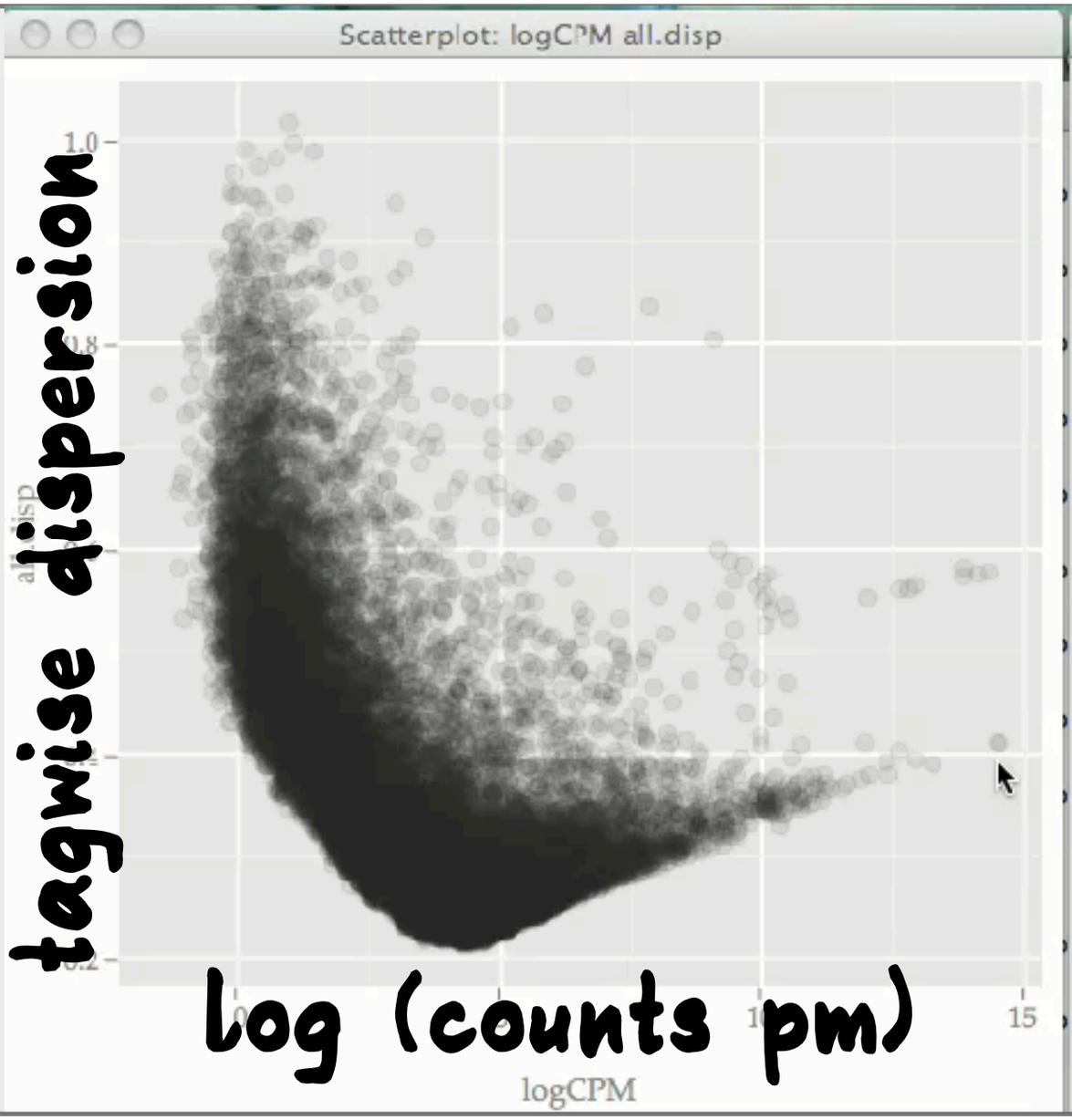
ggplot2

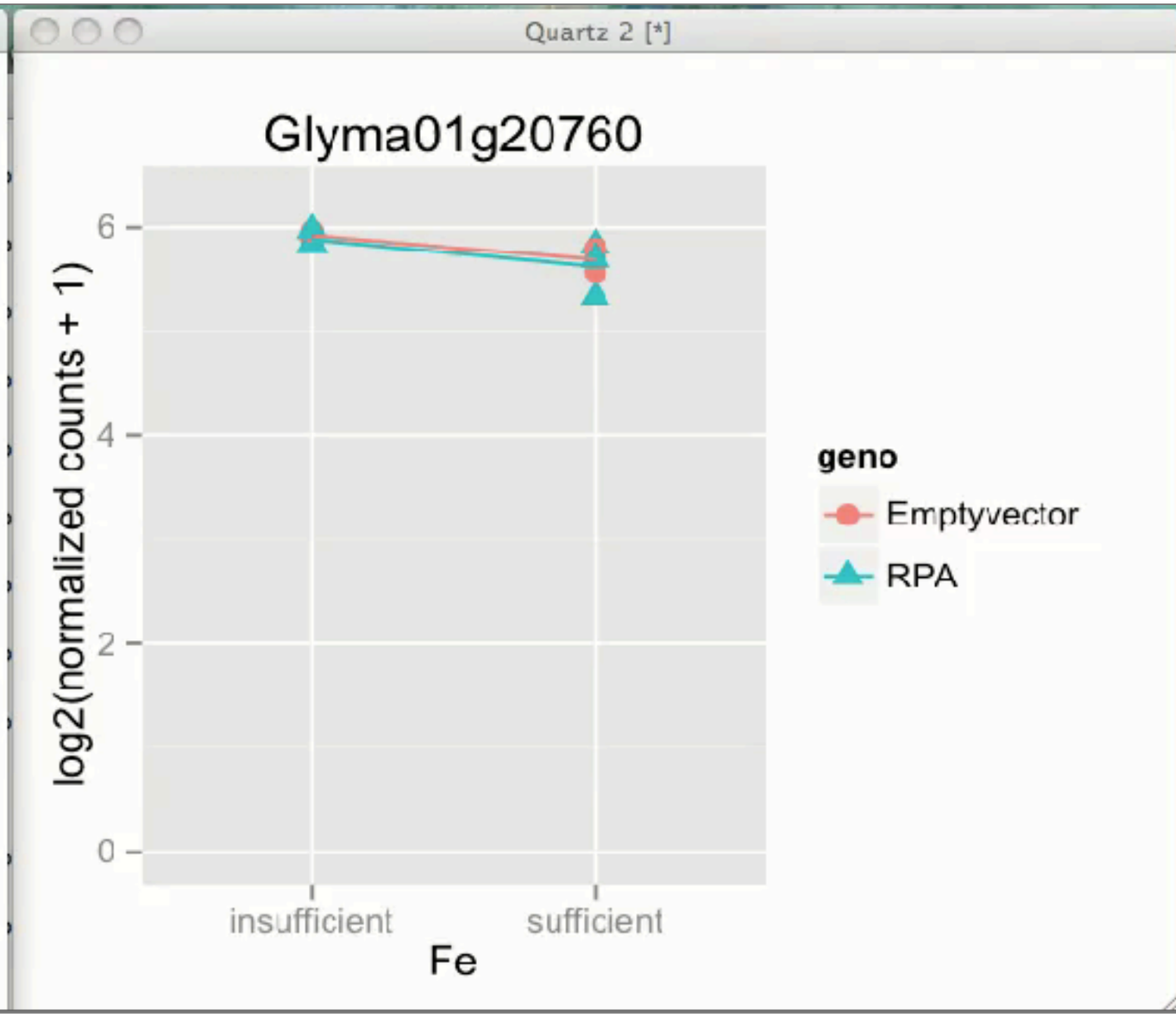




cranvas

ggplot2





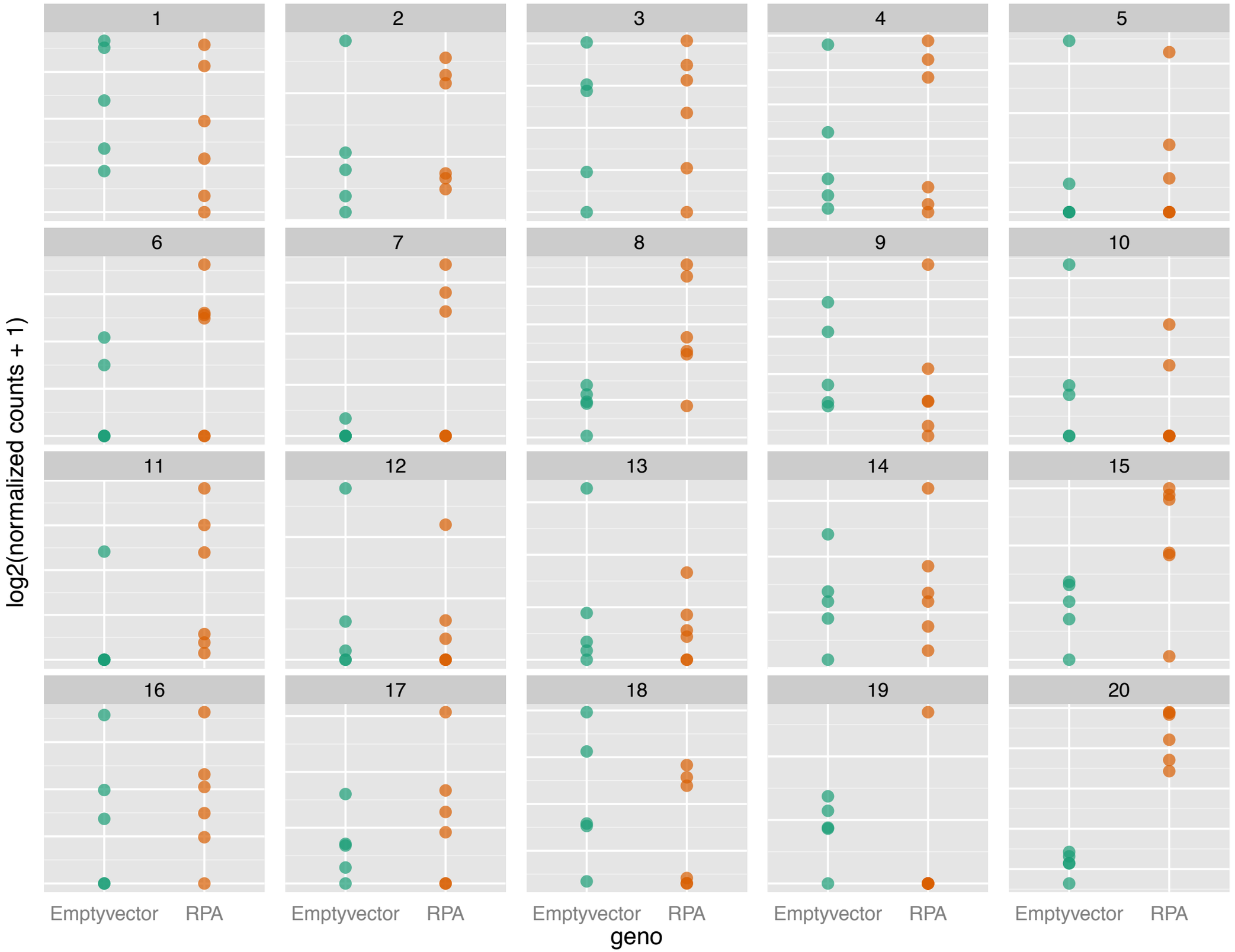
cranvas

ggplot2

So we ran a little experiment

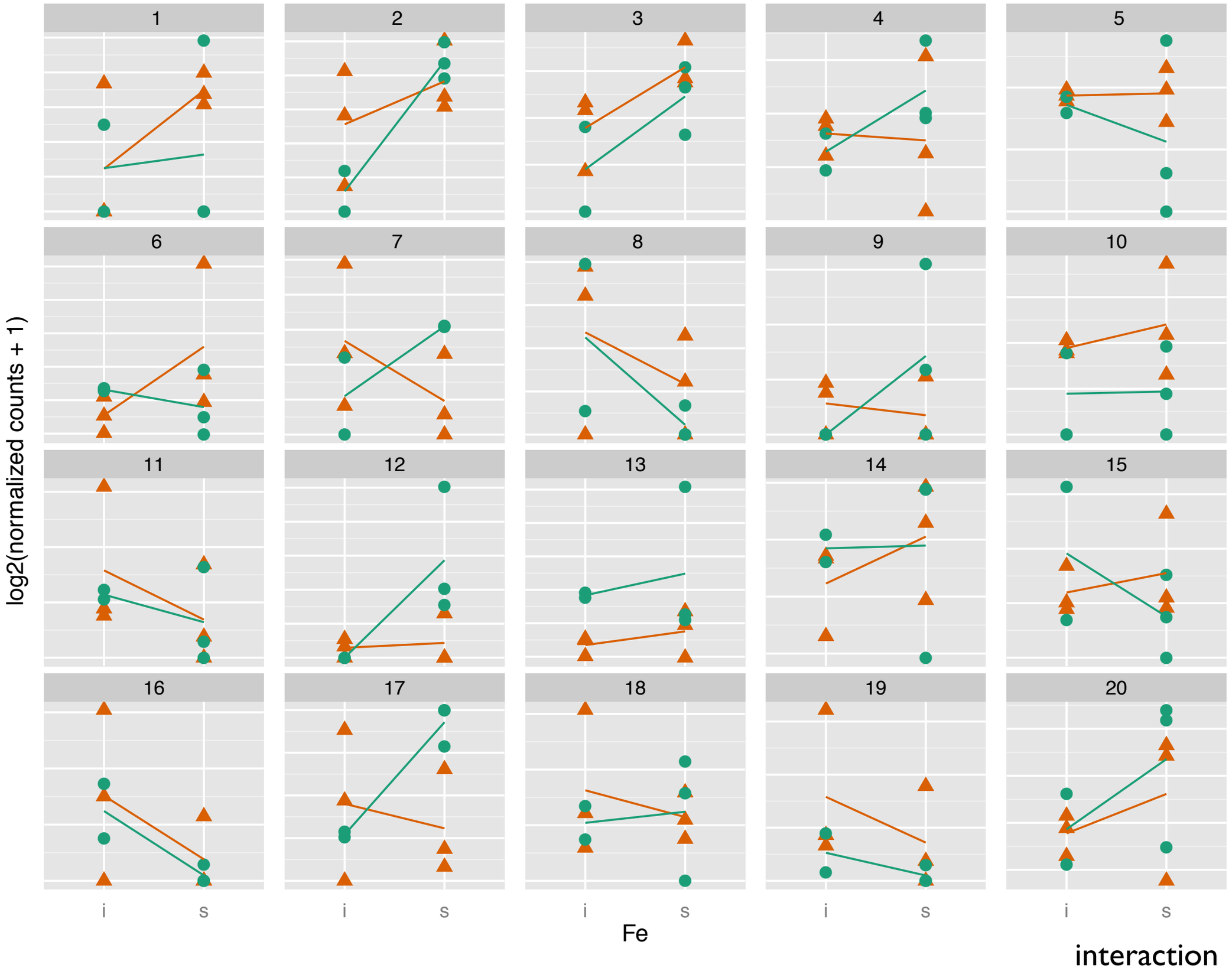
- Compare the results with random results
- Take the experimental design, 2x2x3, and permute the labels
- Re-run the analysis, record most significant gene
- Plot the results

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



geno_1_5, 5/7

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



interaction_2_1, 4/5

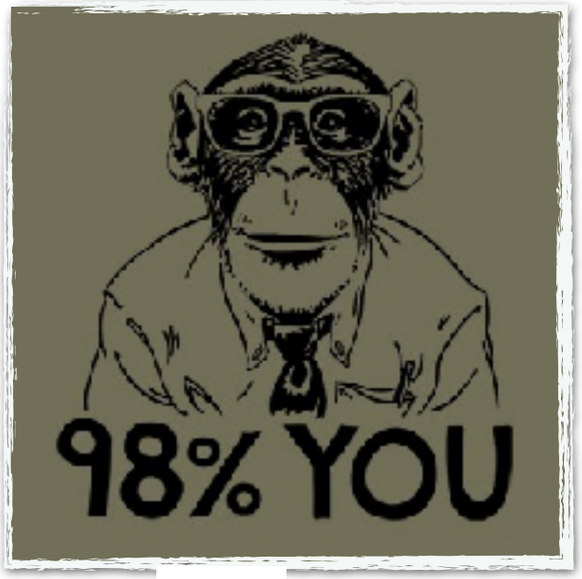
Experiment

- 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!



Human vs chimp

- Data from "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
- Human, chimp (and rhesus) liver RNA
- 3x2(M/F) individuals, 2 reps for each species

Image from son's T-shirt!

Human vs chimp

- Pairwise comparisons of species

$$\log(\mu_g^{s,i}) = \mu_g + \theta_g^s + \delta_g^{sex(i)} + (\theta\delta)_g^{s,sex(i)} + \gamma_g^i$$

$$M_0 : \mu_g \neq 0, \theta_g^s = 0, \delta_g^{sex(i)} \neq 0, (\theta\delta)_g^{s,sex(i)} = 0$$

$$M_1 : \mu_g \neq 0, \theta_g^s \neq 0, \delta_g^{sex(i)} \neq 0, (\theta\delta)_g^{s,sex(i)} = 0$$

Likelihoods compared, FDR<0.05

Table 1. Numbers of differentially expressed genes between species at FDR < 0.05

	All	Males only	Females only
Human–chimpanzee	3335	1787	1037
Human–rhesus	6030	3002	3493
Chimpanzee–rhesus	5549	3109	3088

Human vs chimp

- Re-analyzed using edgeR, exactTest
- (Yes, not taking dependencies into account - but a quick re-do of analysis wanted)
- Just Human-Chimp
- Yields 3630 differentially expressed genes, at FDR<0.01, mostly overlapping with published results

Visual testing

- Create multiple sets of permutations of the labels of human, chimp
- Conduct edgeR/exactTest on each of the permutations
- Record the top 2500 genes based on p-value
- Make lineups of j'th ordered gene of actual data against those of permuted data

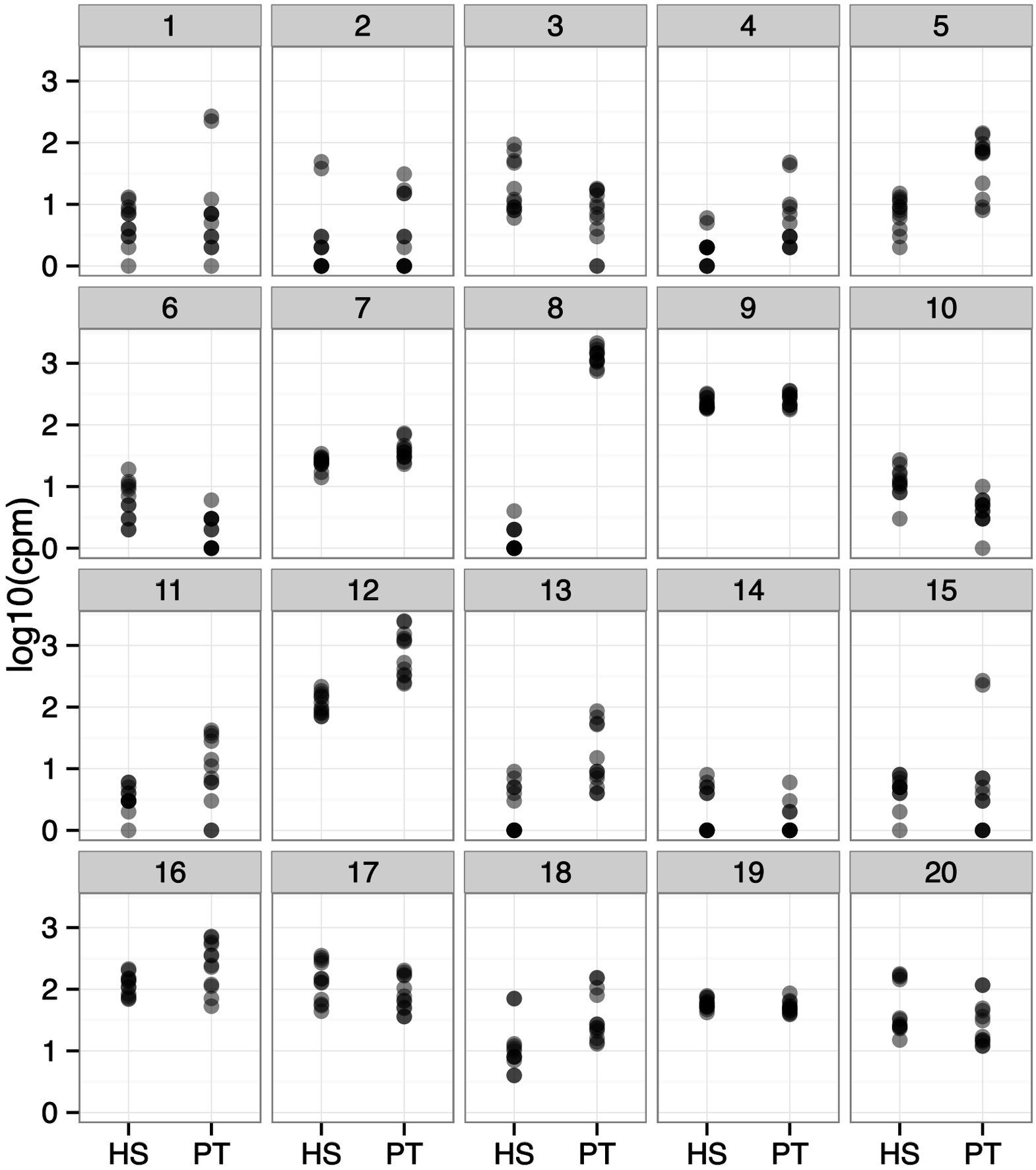
You try

- Pick one plot among the 20
- "Which plot has the largest vertical difference between the two groups?"

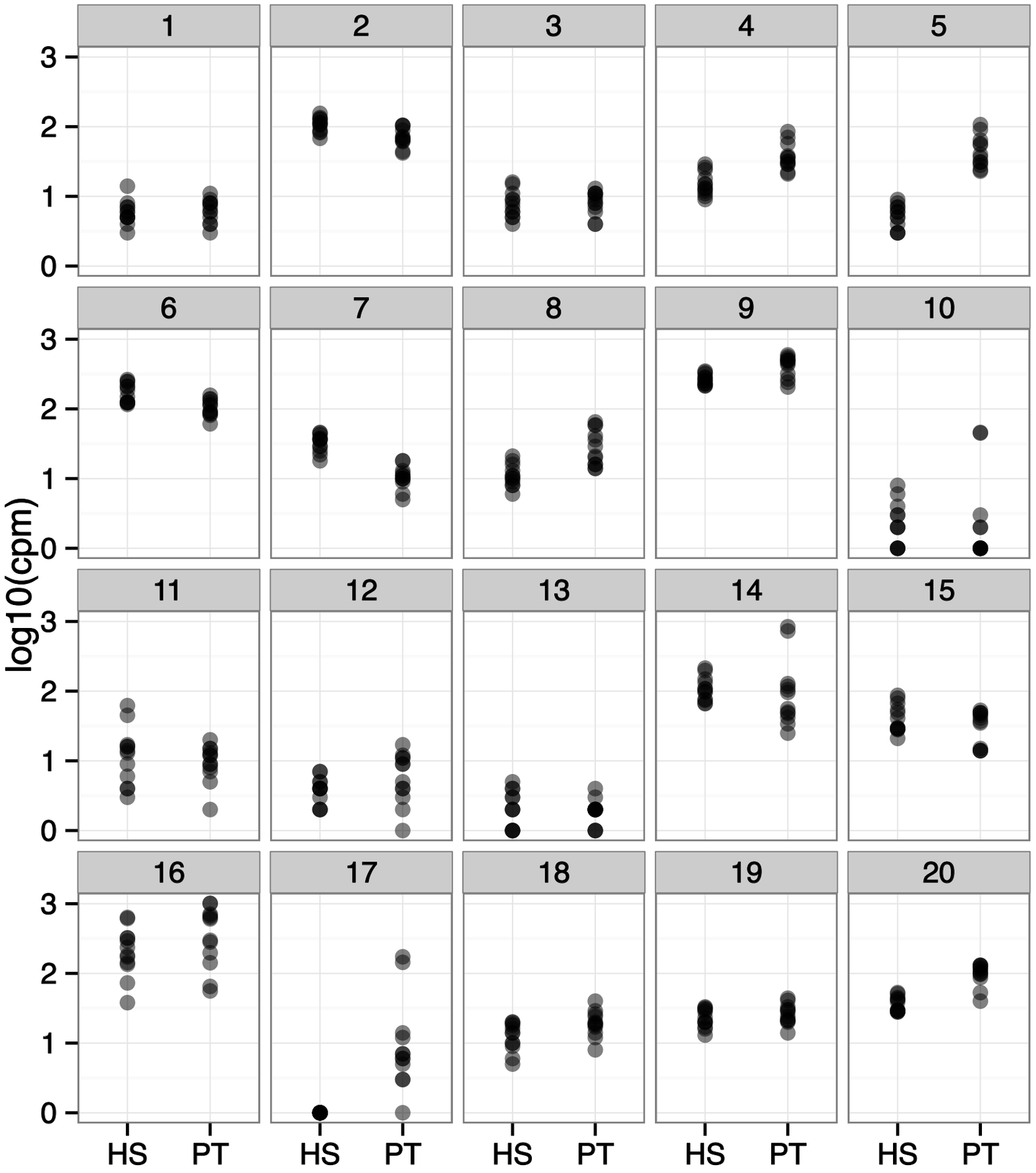
Point your mobile device to this web page

goo.gl/gG60uR

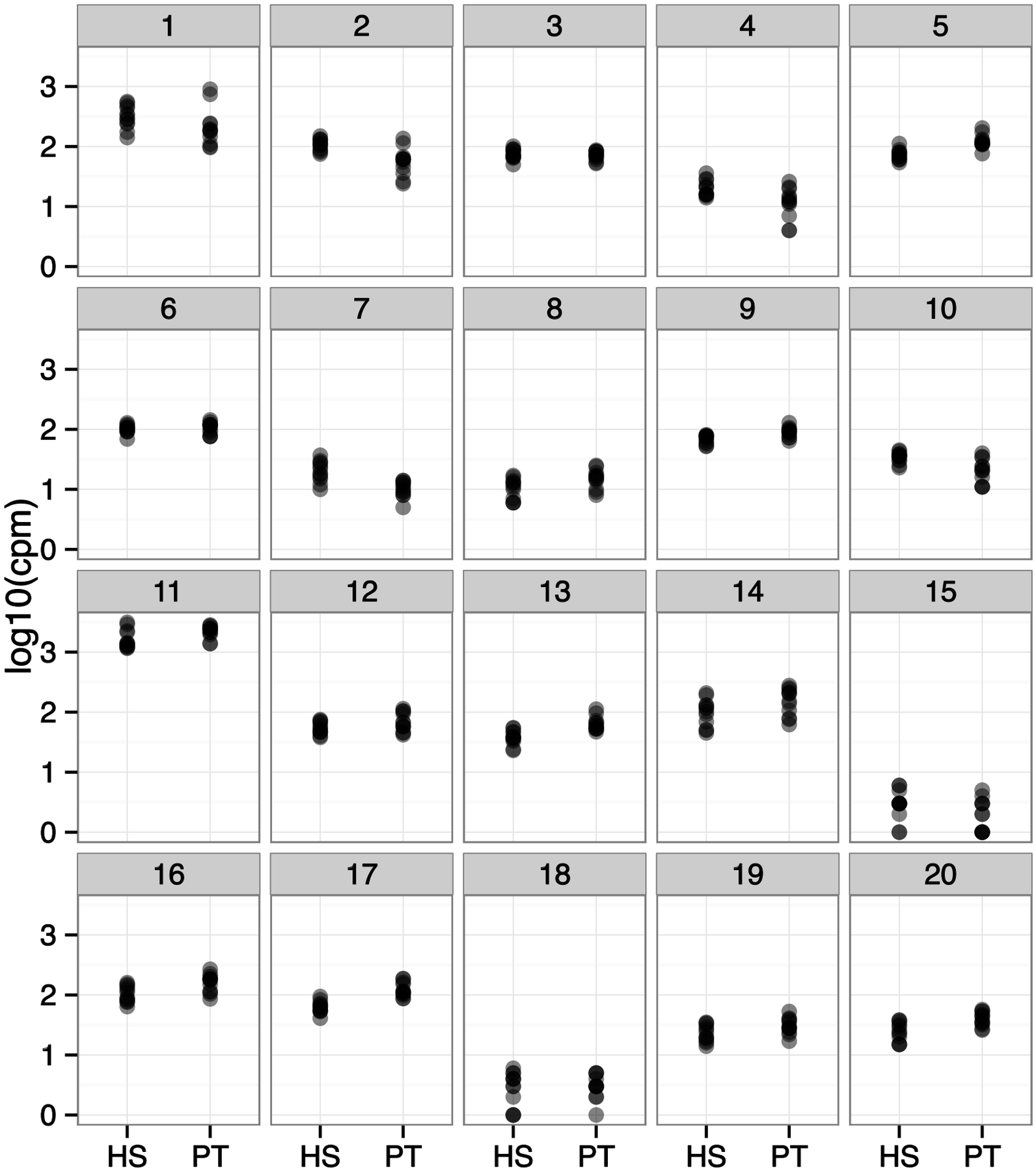
Human-chimp 1



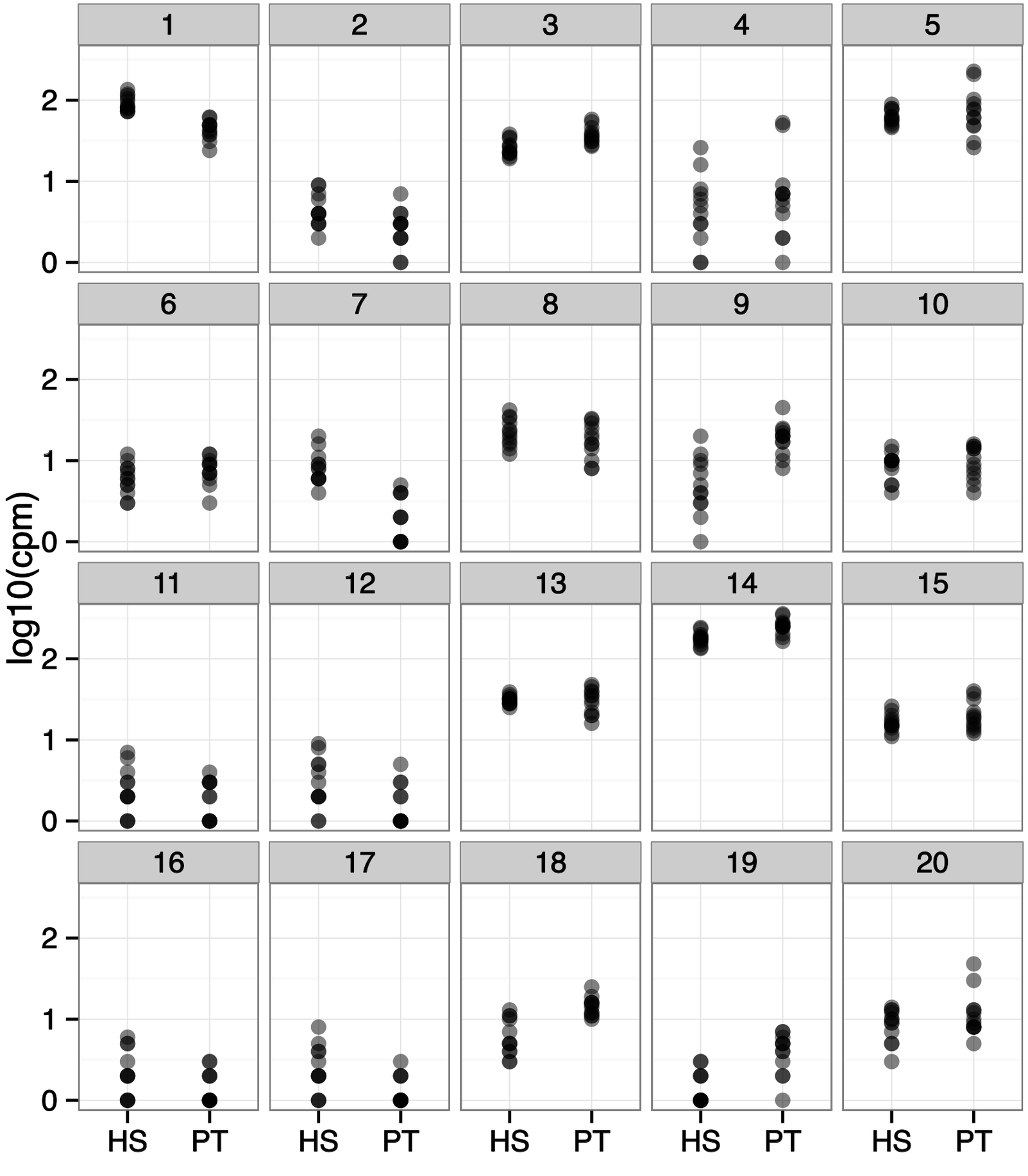
Human-chimp 2



Human-chimp 3



Human-chimp 4



Actual data is in positions 8, 5, 17, 18

Turk study

- Lineups of the 2-11'th, 95-104'th, 995-1004'th, 1995-2004'th ordered genes
- Two replicates of lineups made with different nulls, and different positions of actual data
- Turkers evaluated blocks of 10 randomly selected lineups
- Combine results from turkers

http://www.unomaha.edu/mahbubulmajumder/html/experiments.html

Significance

- If there is no difference in gene expression the chance of one person detecting the actual plot out of 20 is 1/20=0.05
- Multiple people follows:

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

Significance

- If there is no difference in gene expression the chance of one person detecting the actual plot out of 20 is 1/20=0.05
- Multiple people follows:

Number of independent observers

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

Significance

- If there is no difference in gene expression the chance of one person detecting the actual plot out of 20 is 1/20=0.05
- Multiple people follows:

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

Significance

- If there is no difference in gene expression the chance of one person detecting the actual plot out of 20 is 1/20=0.05
- Multiple people follows:

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

Number of observers choosing data plot

Significance

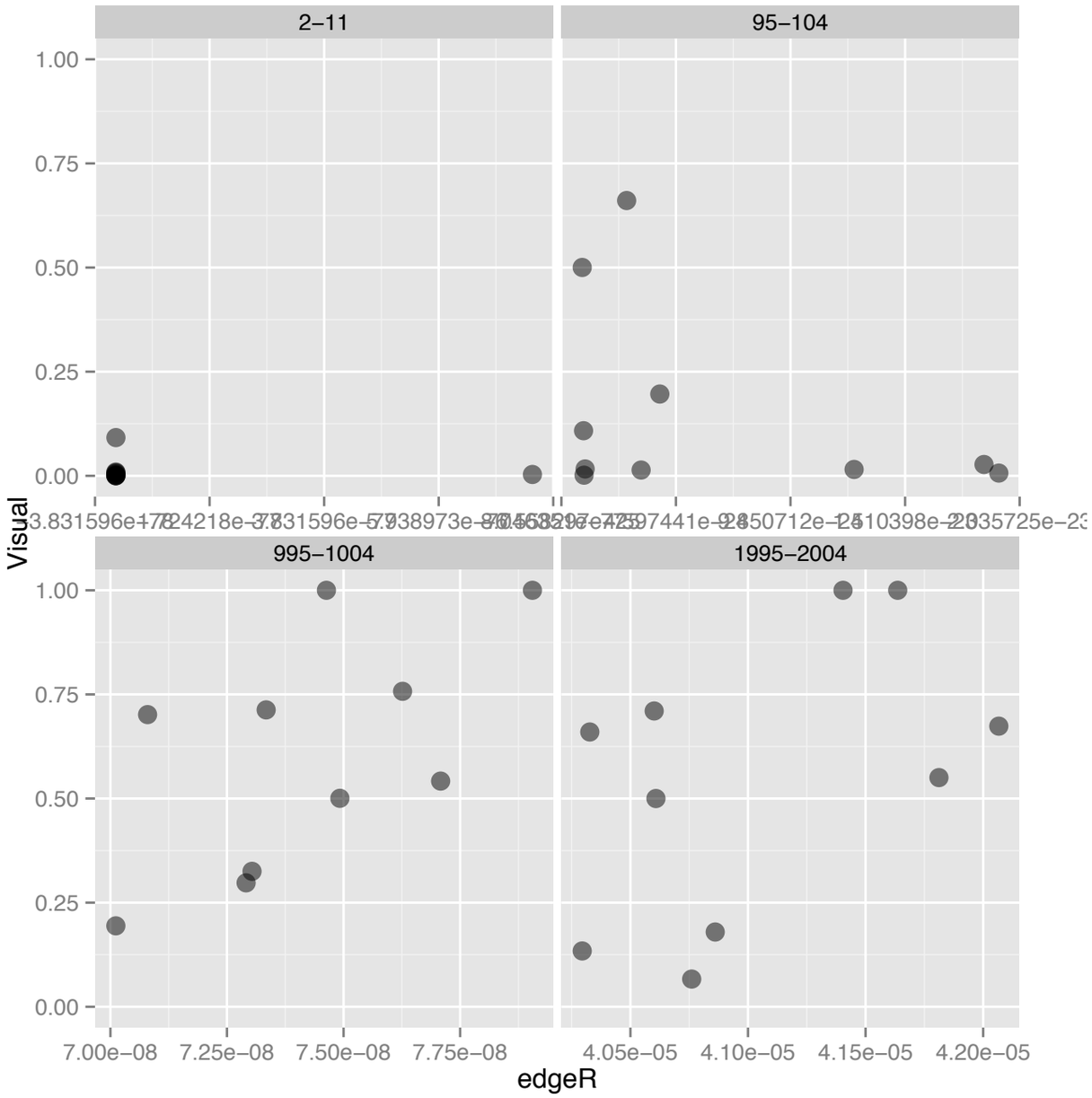
- If there is no difference in gene expression the chance of one person detecting the actual plot out of 20 is 1/20=0.05
- Multiple people follows:

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

Results

- Genes in the top 10 have a clear difference
- From 900's down difference is consistent with randomness

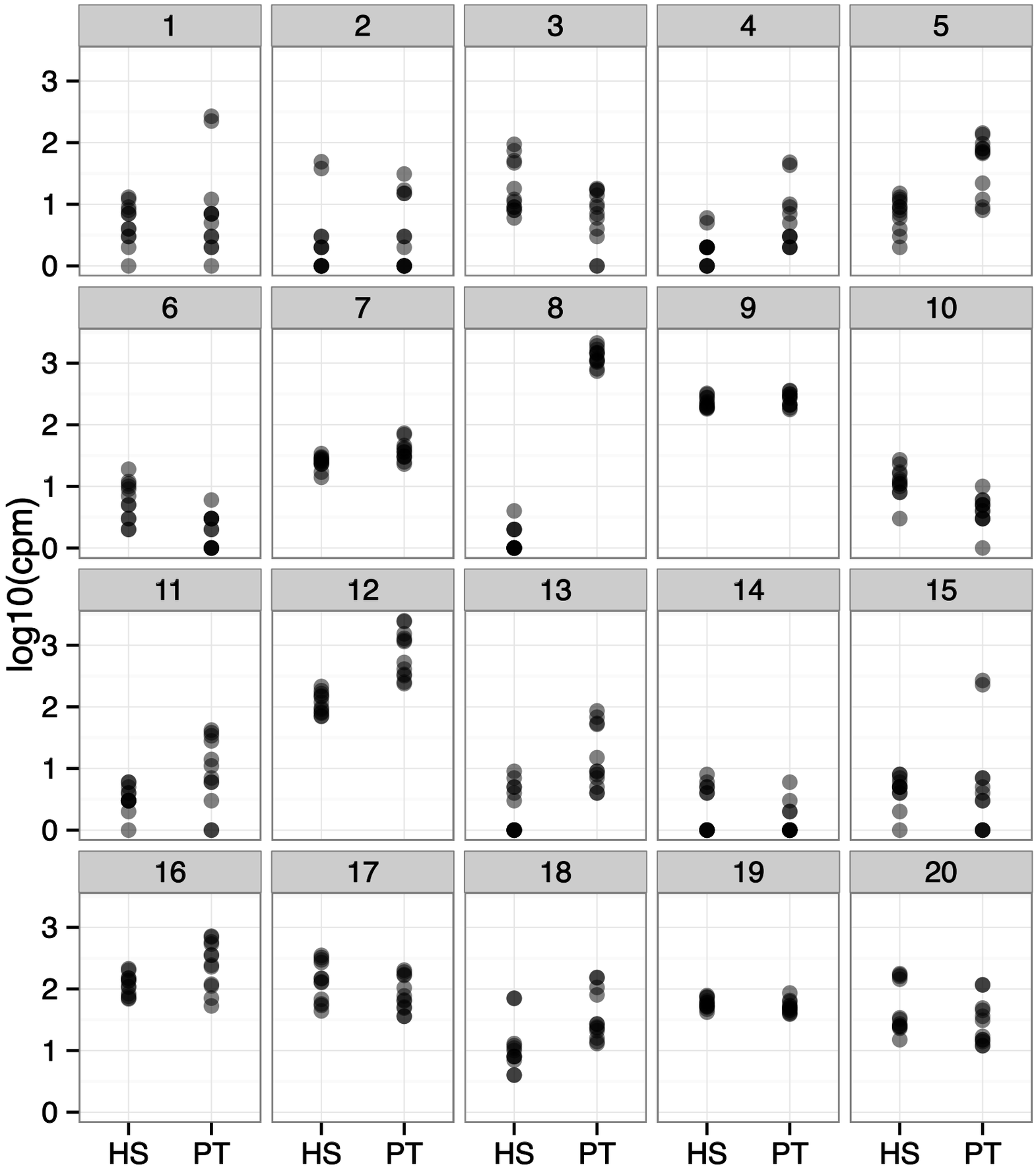
Results



- Visual p-value 0 when edgeR p-value is really tiny, top 20
- Positive association with p-values with order ~1000th, ~2000th

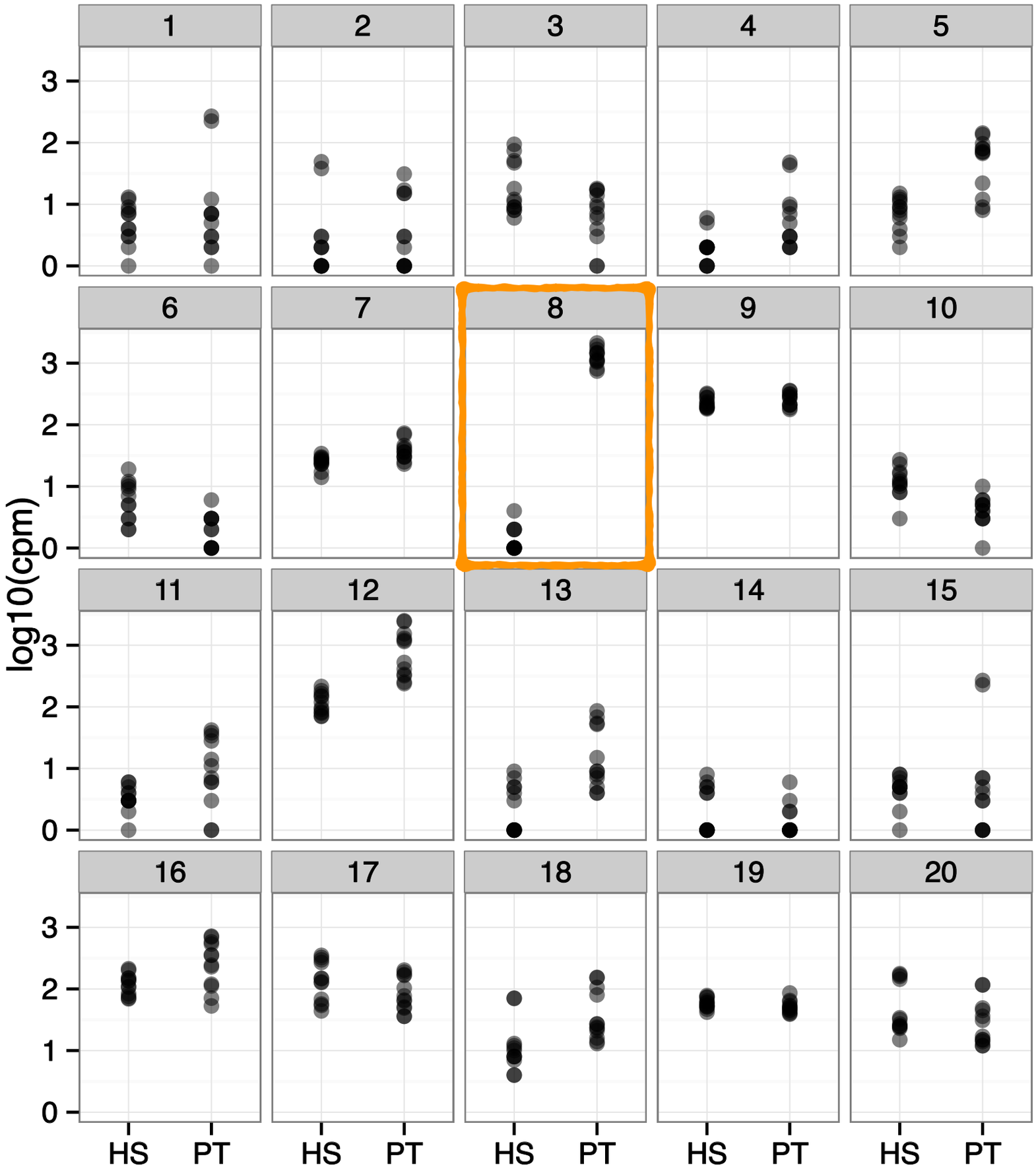
Human-chimp 1

2nd



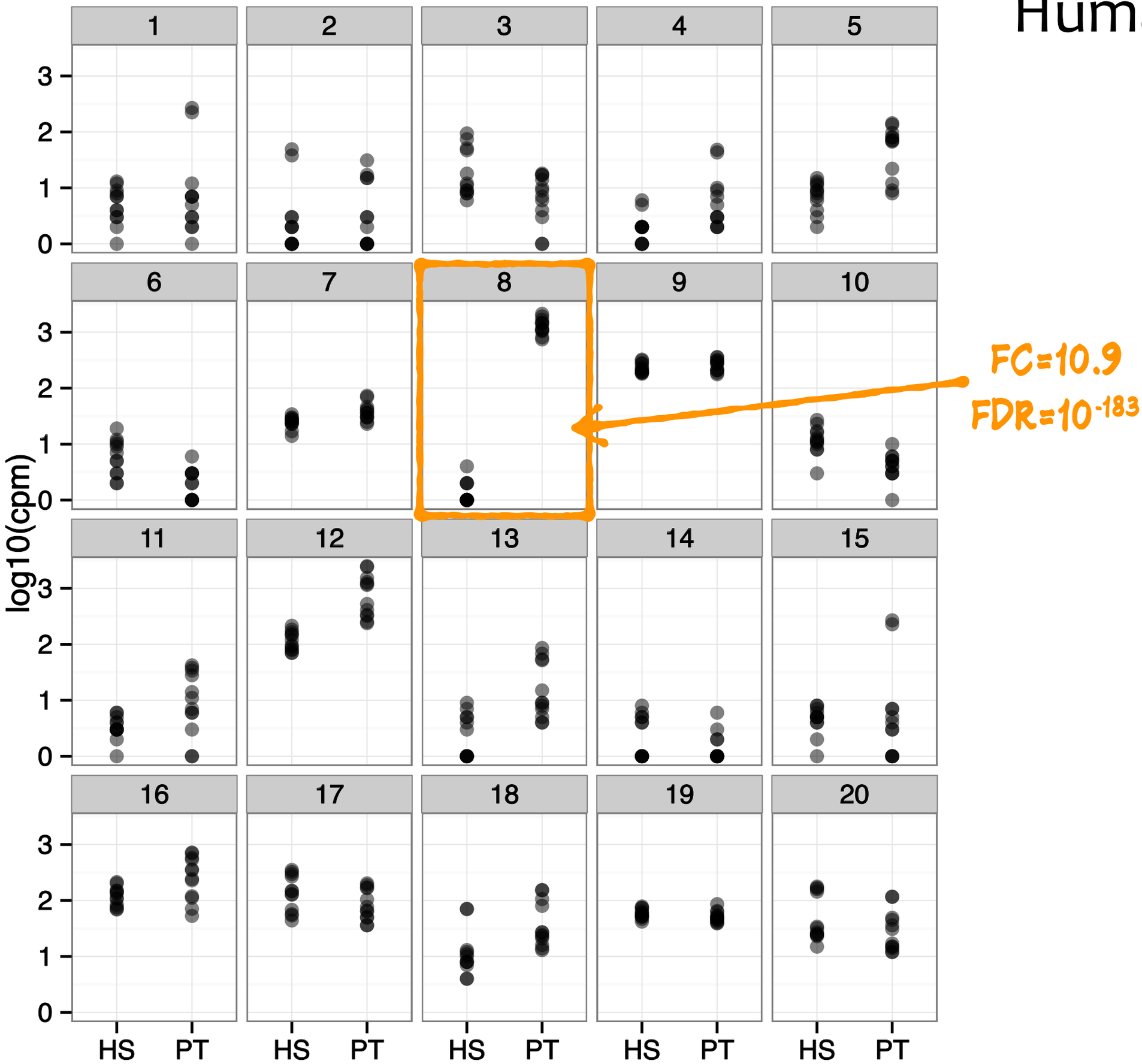
Human-chimp 1

2nd



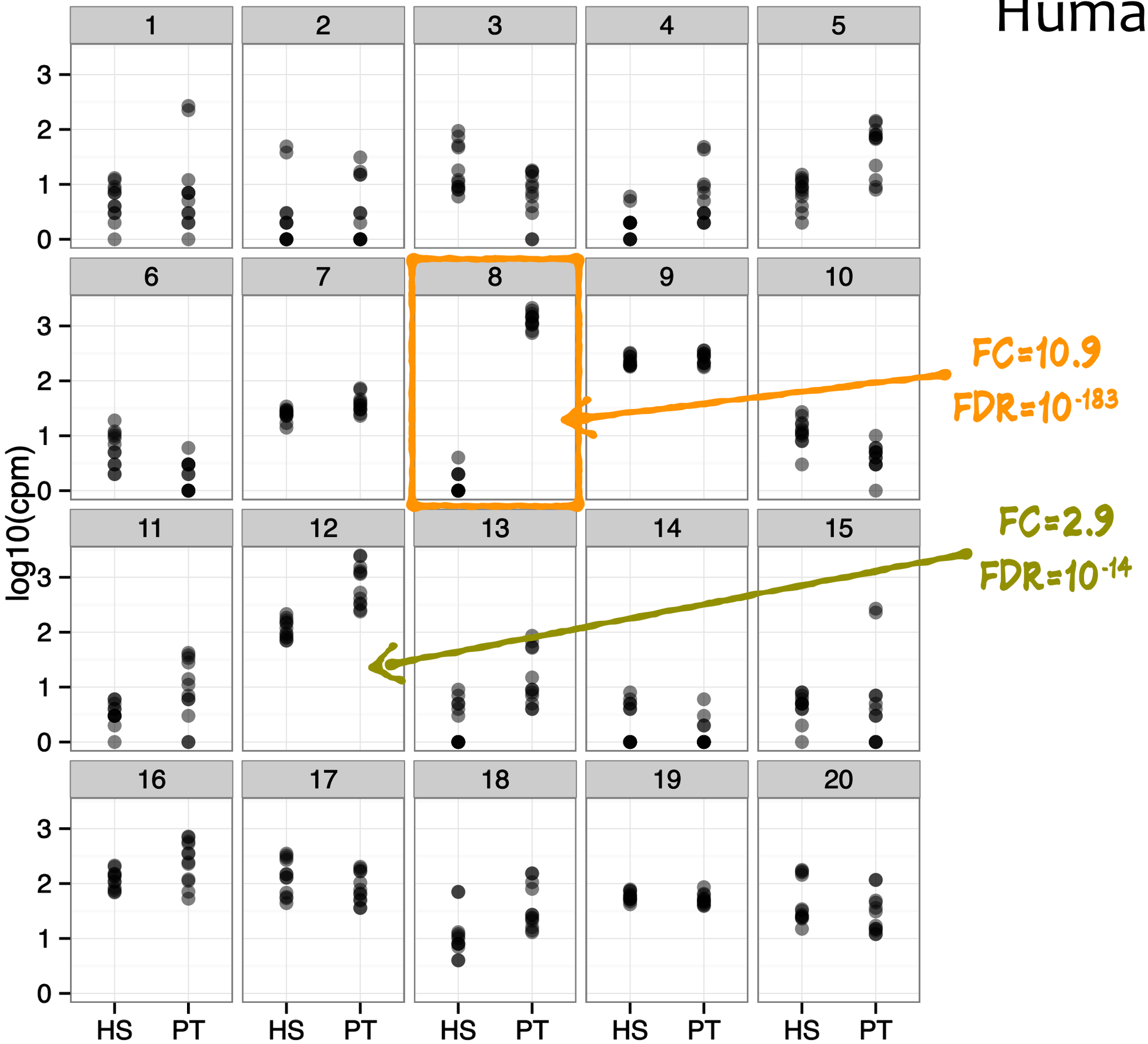
Human-chimp 1

2nd



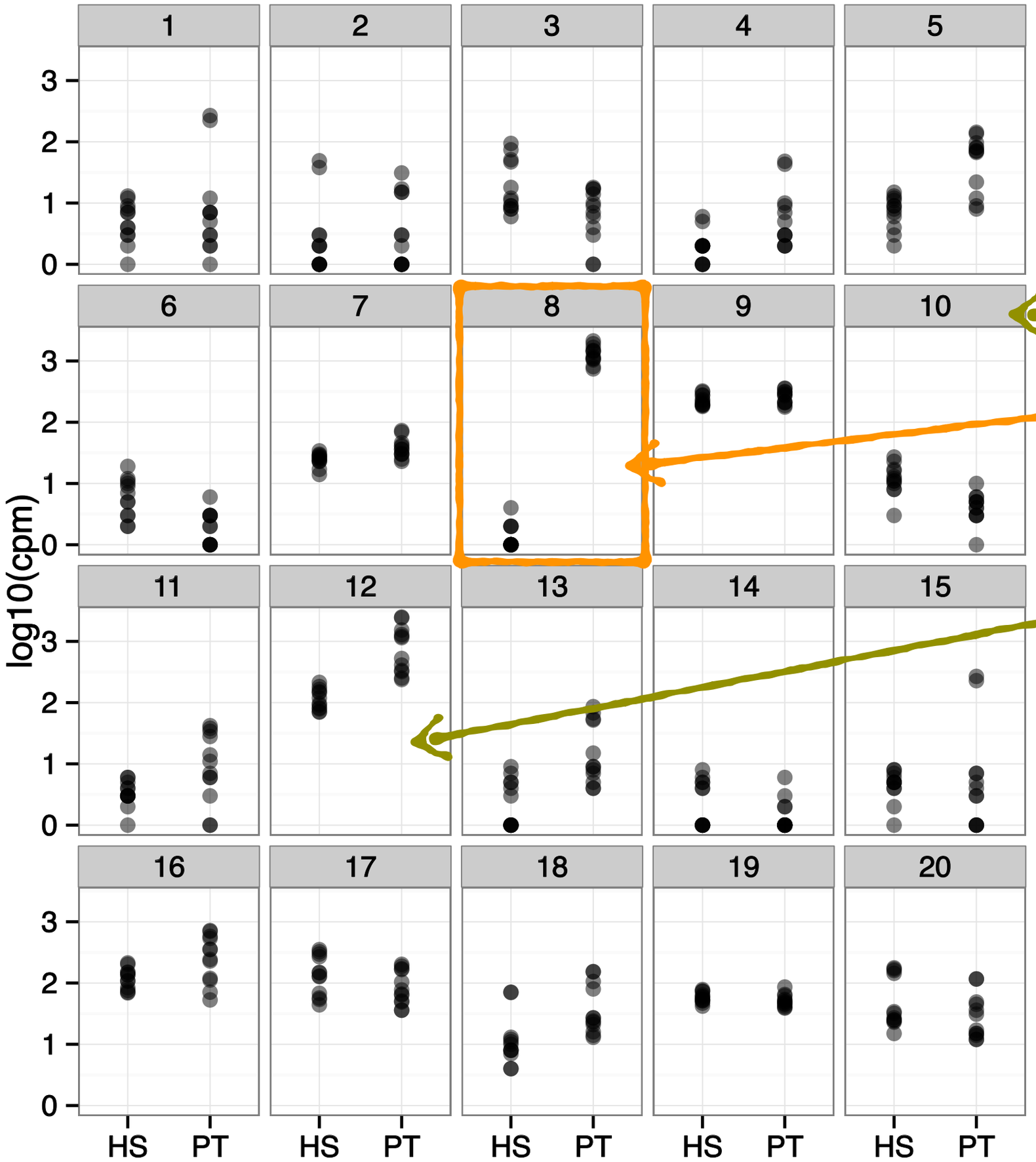
Human-chimp 1

2nd



Human-chimp 1

2nd



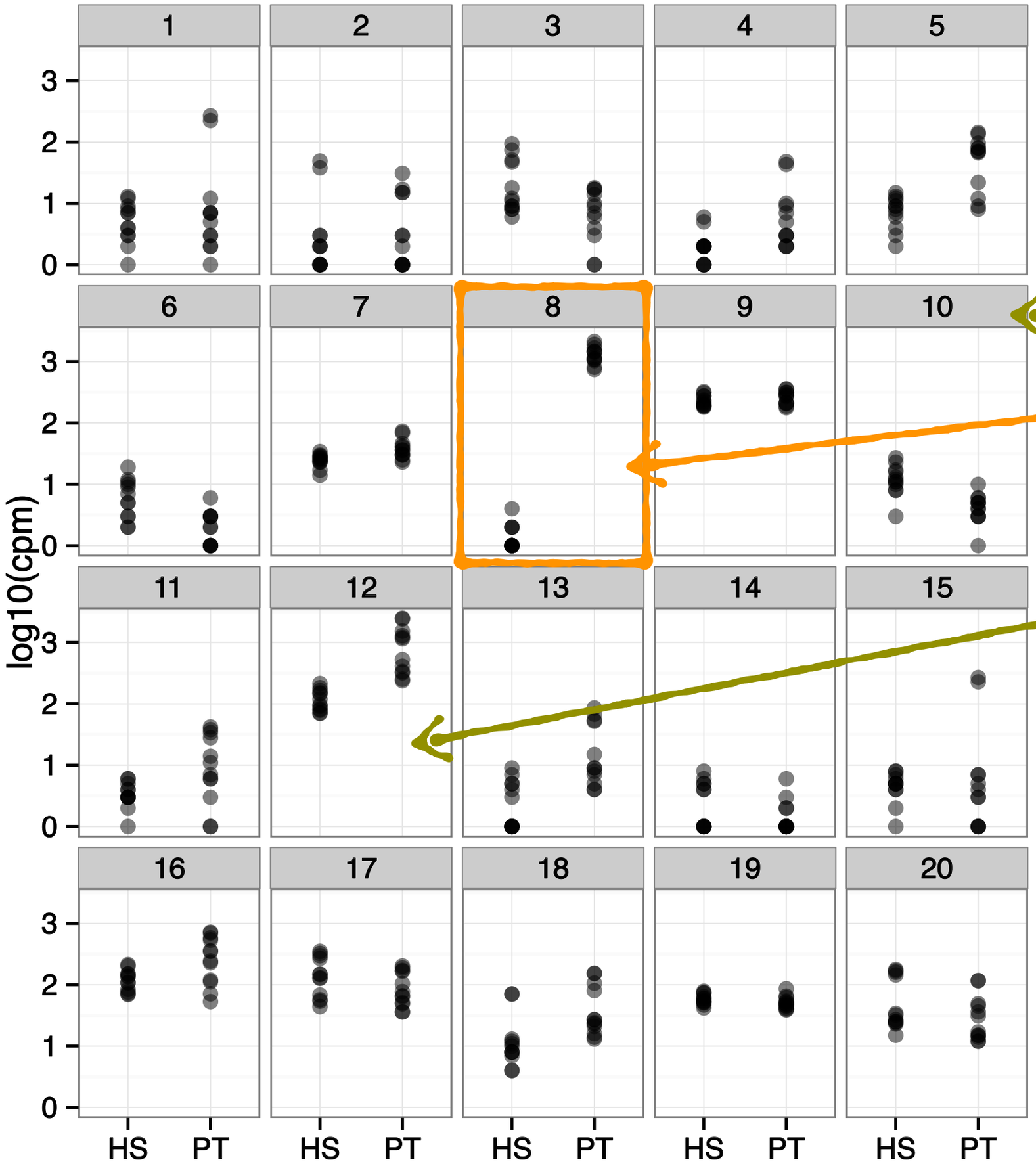
FC=3.0
FDR=10^{-11}

FC=10.9
FDR=10^{-183}

FC=2.9
FDR=10^{-14}

Human-chimp 1

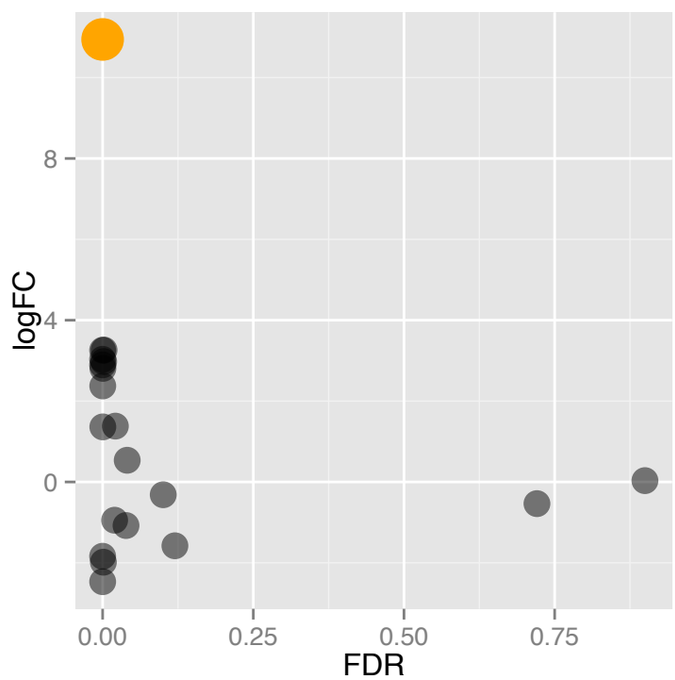
2nd



FC=3.0
FDR=10^{-11}

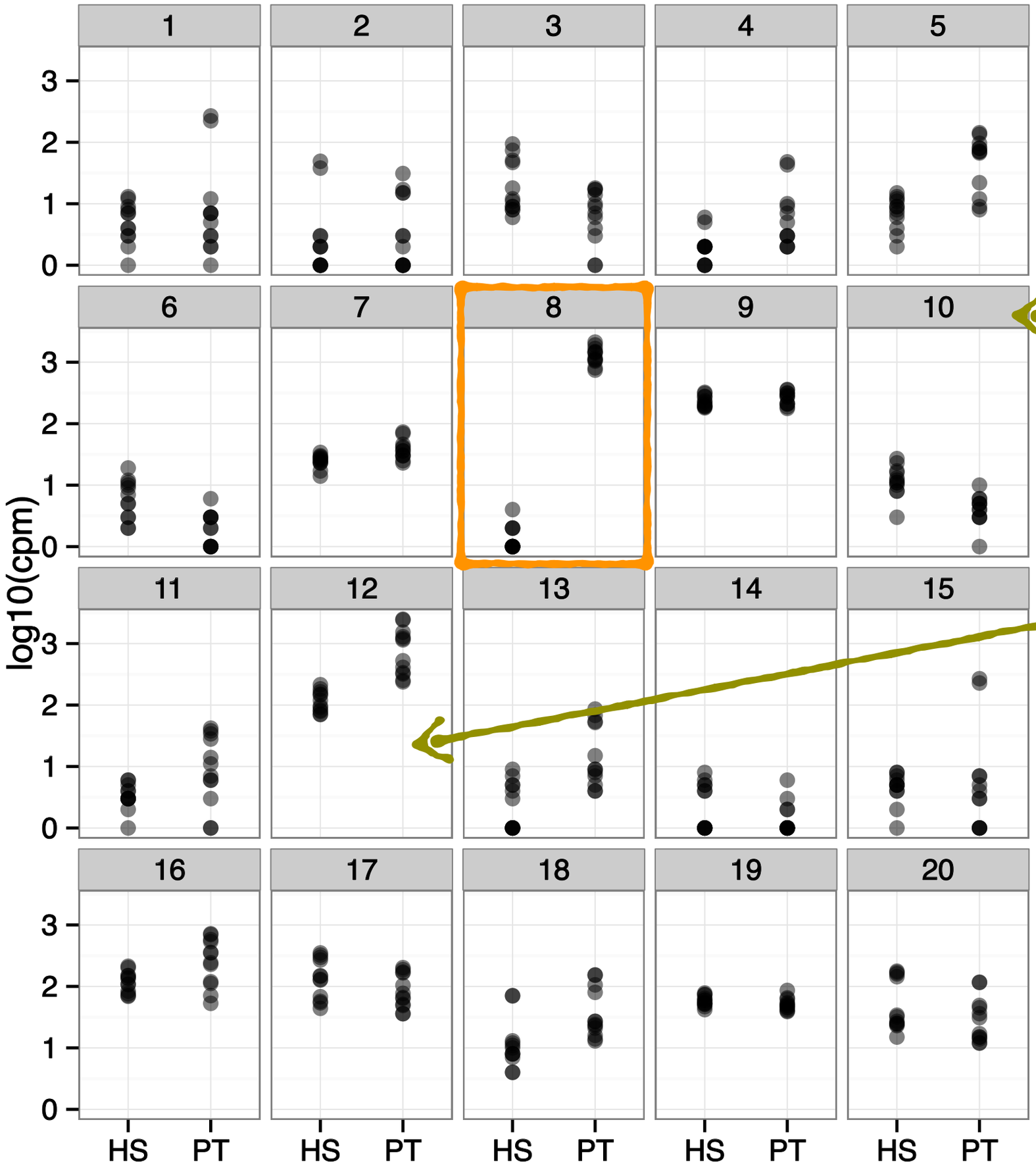
FC=10.9
FDR=10^{-183}

FC=2.9
FDR=10^{-14}



Human-chimp 1

2nd

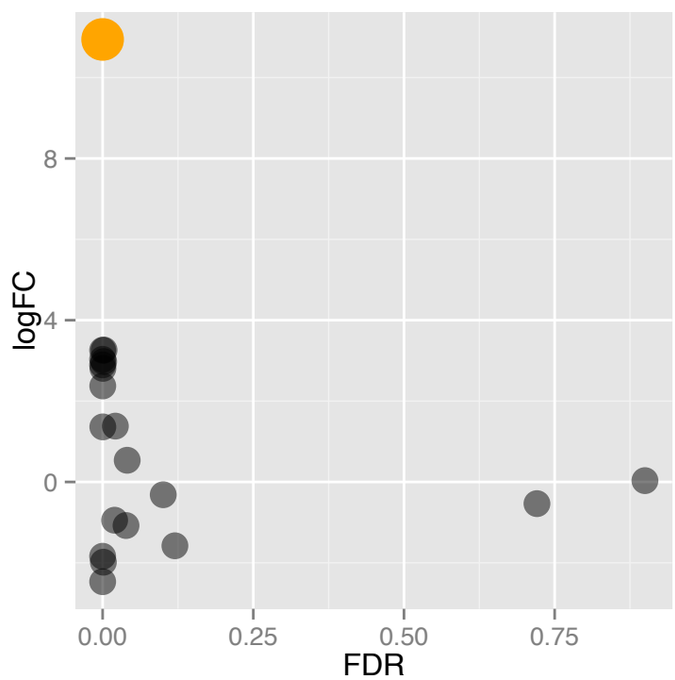


FC=3.0

FDR=10^{-11}

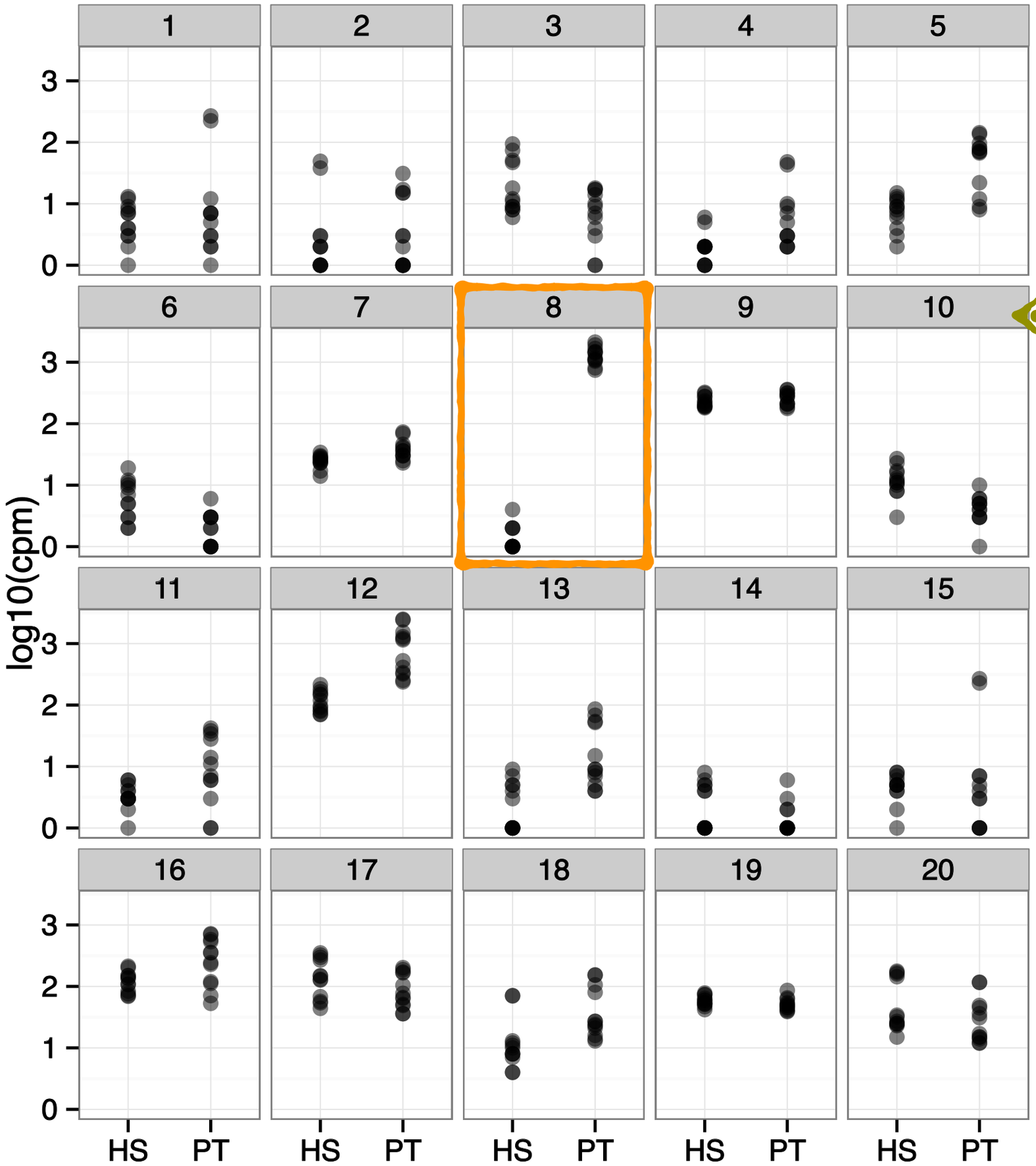
FC=2.9

FDR=10^{-14}



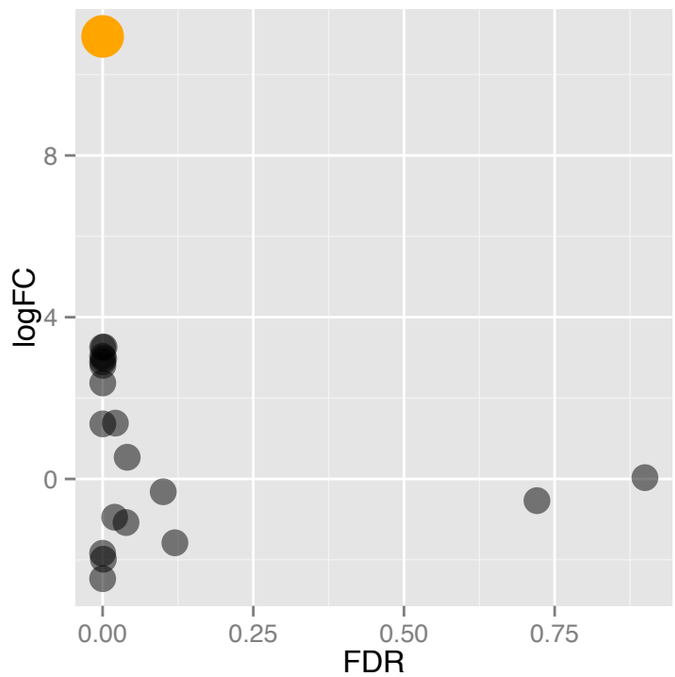
Human-chimp 1

2nd



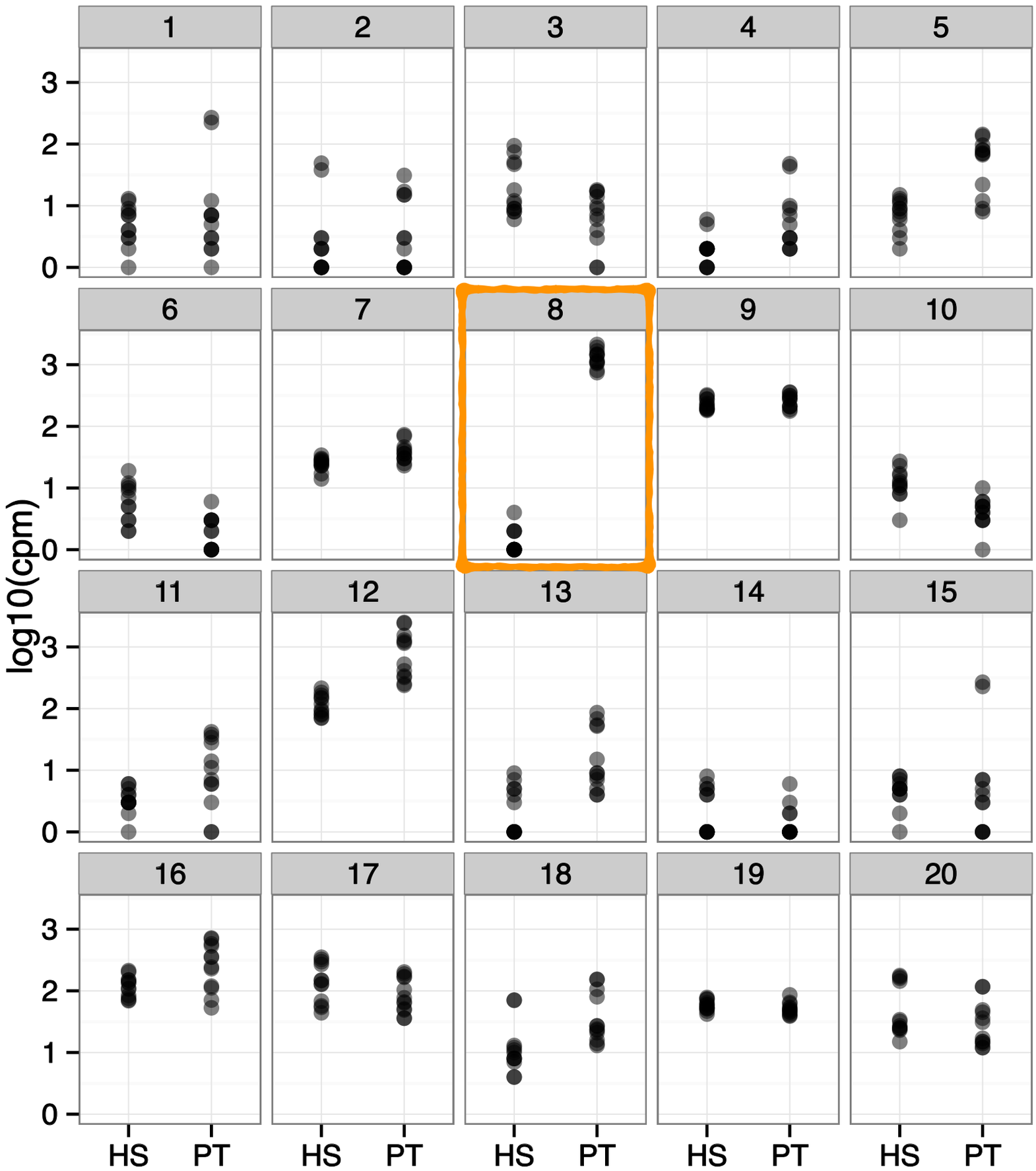
FC=3.0

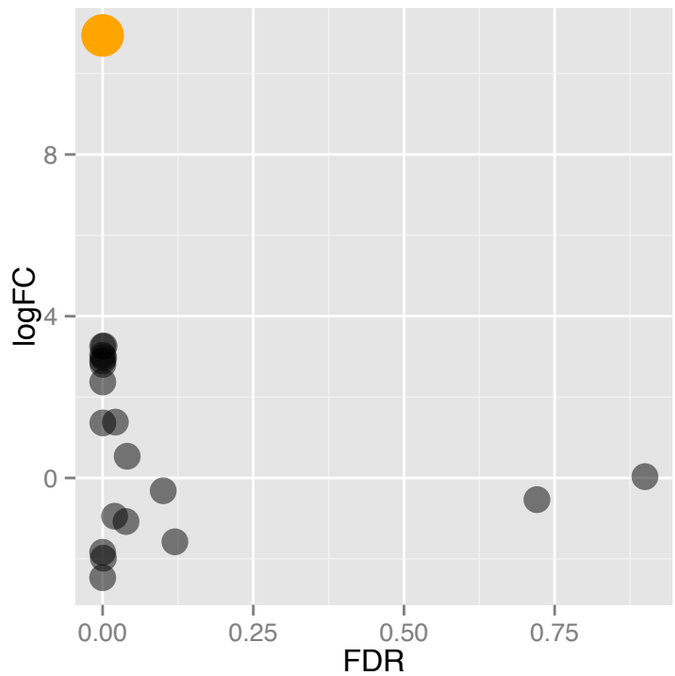
FDR=10^{-11}



Human-chimp 1

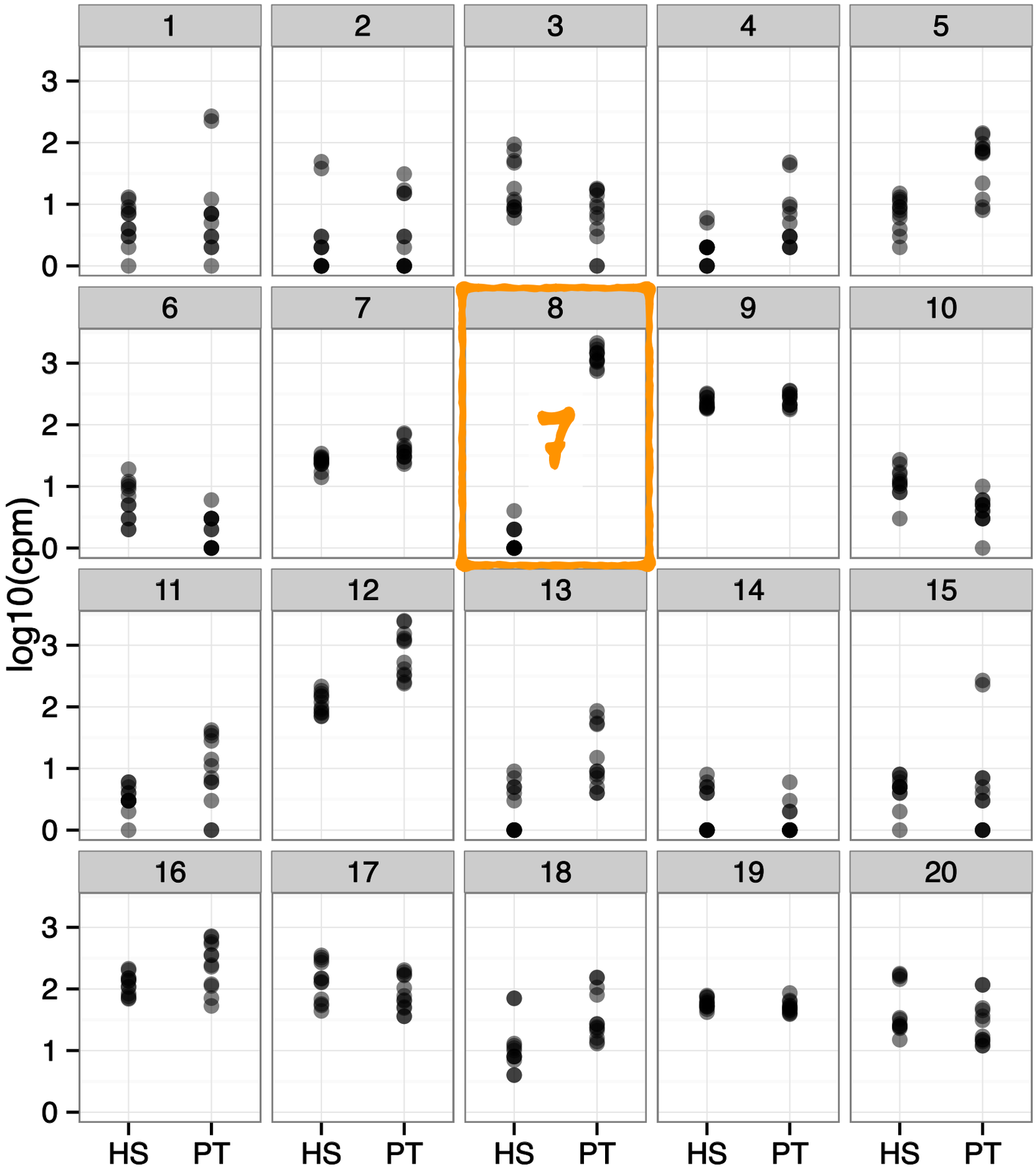
2nd

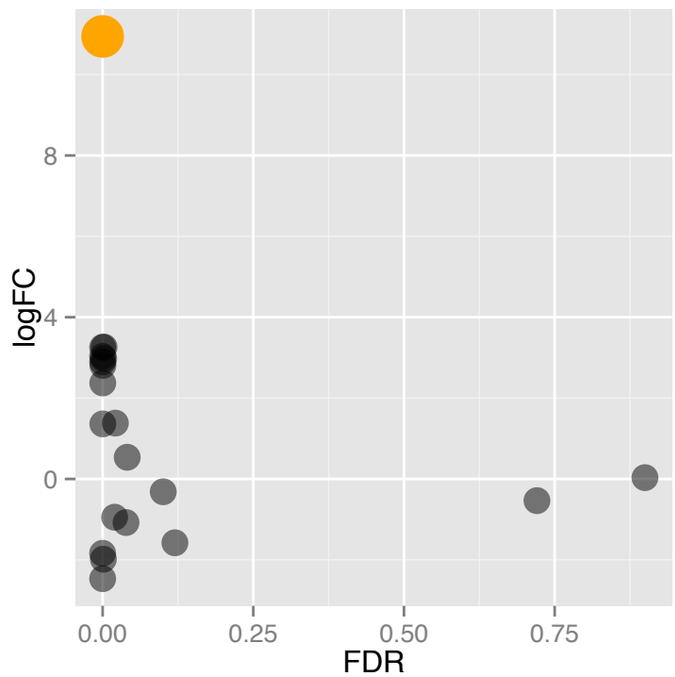




Human-chimp 1

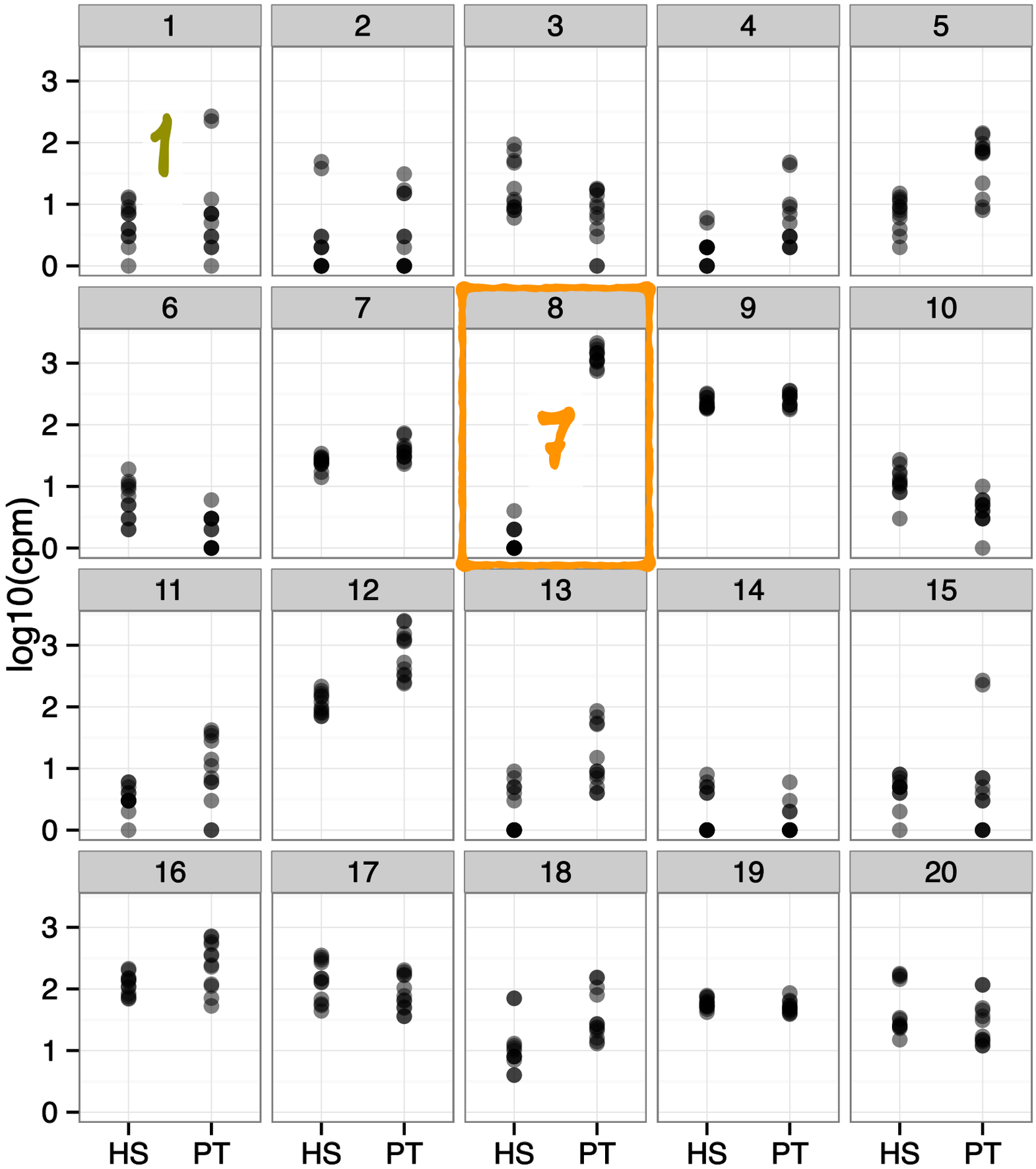
2nd

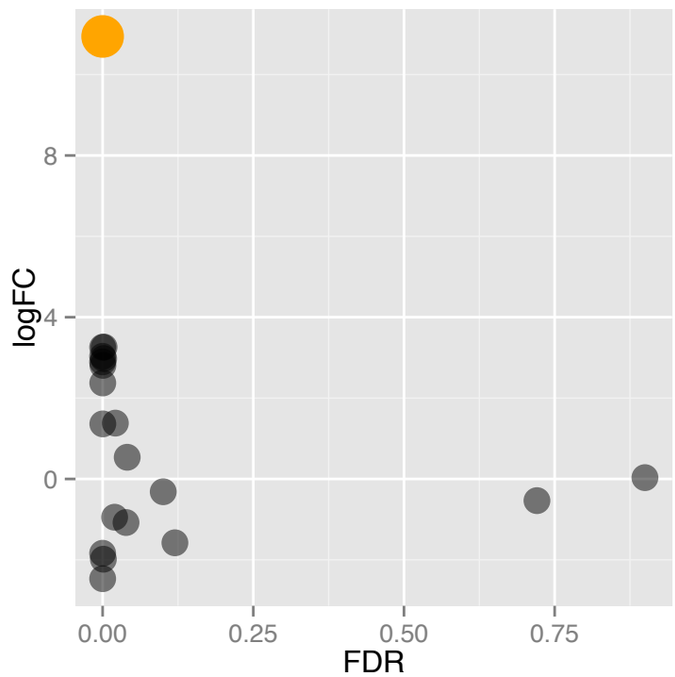




Human-chimp 1

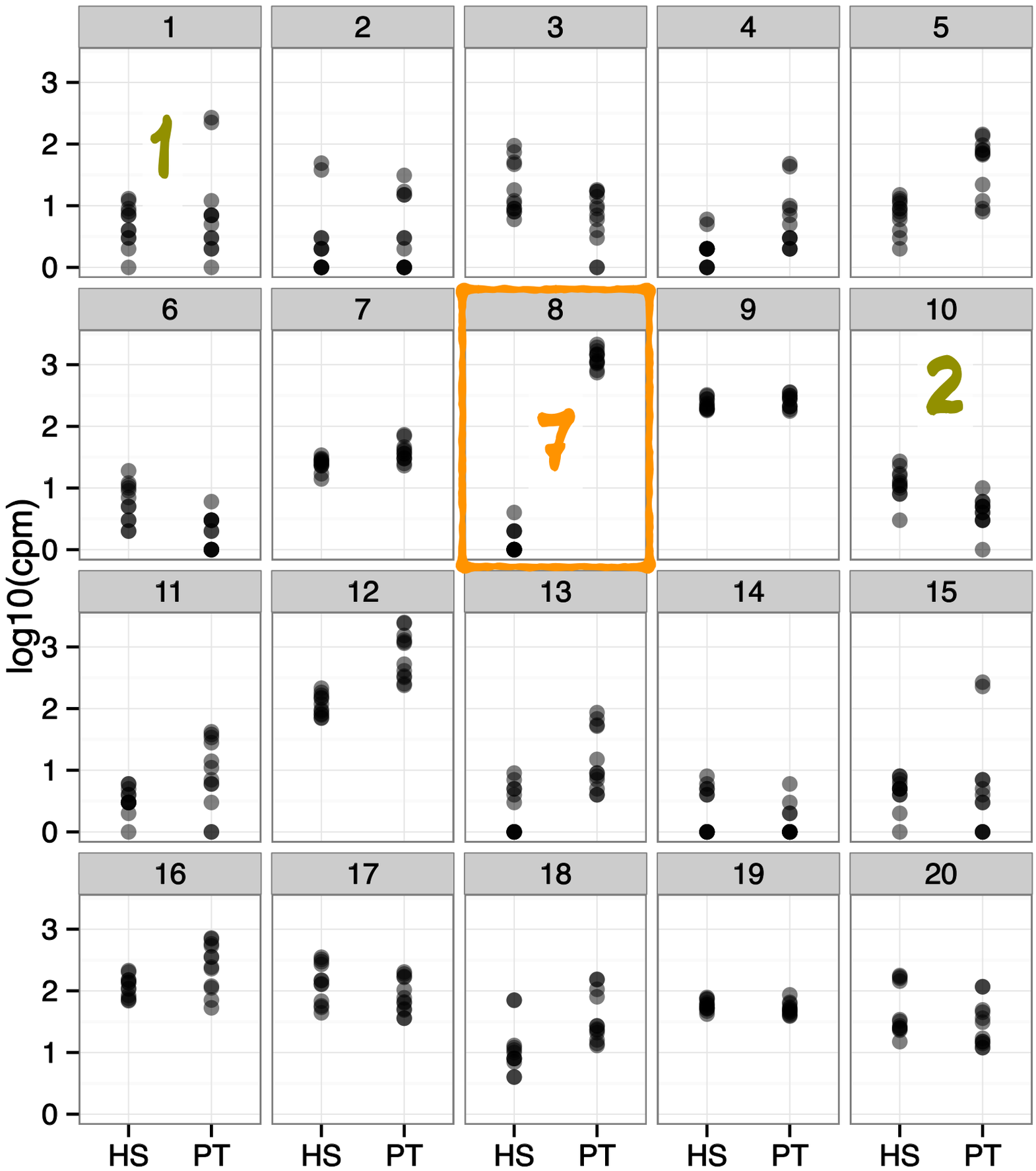
2nd

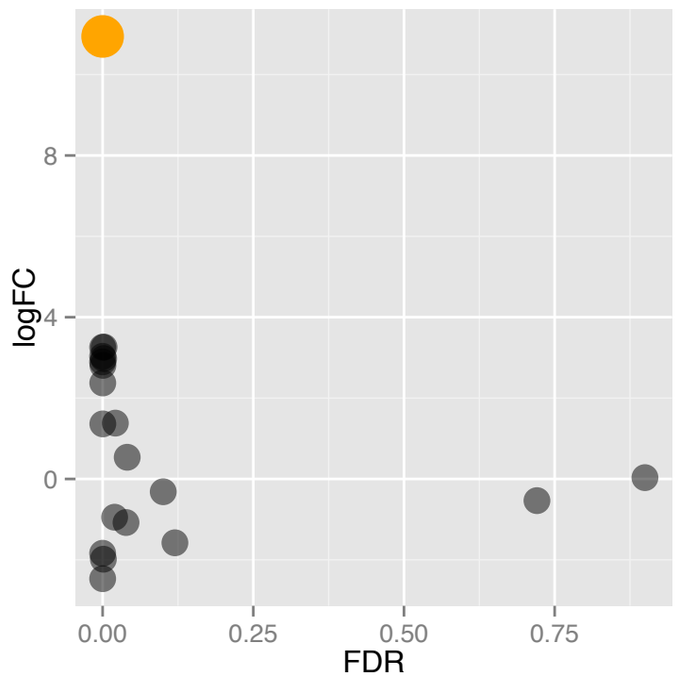




Human-chimp 1

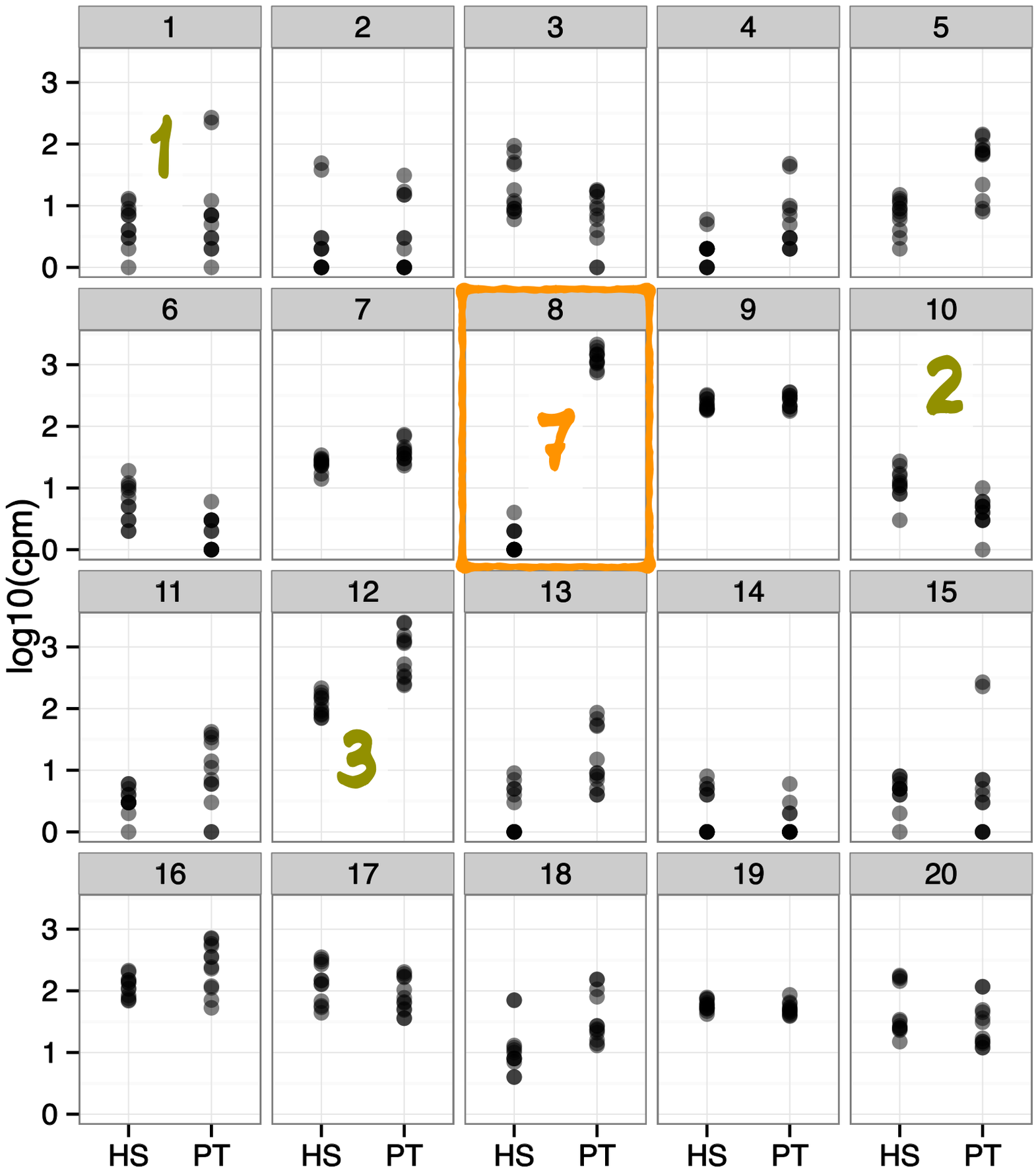
2nd

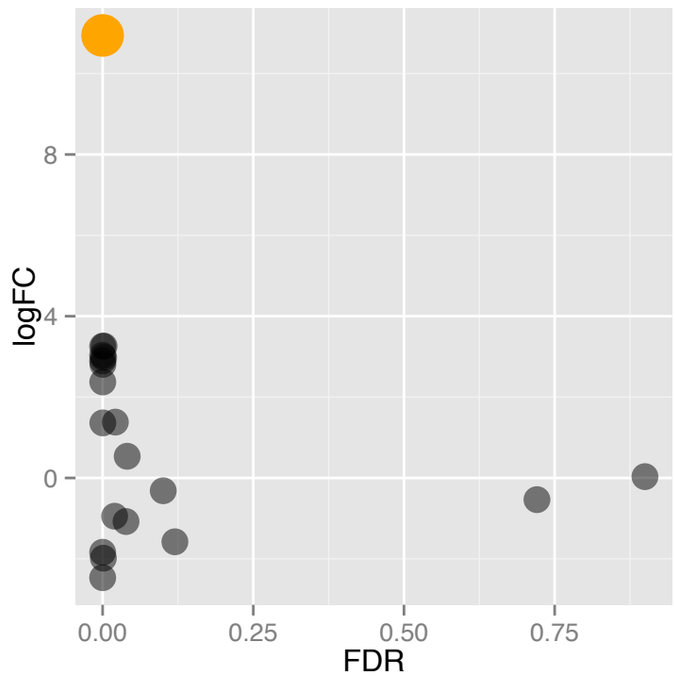




Human-chimp 1

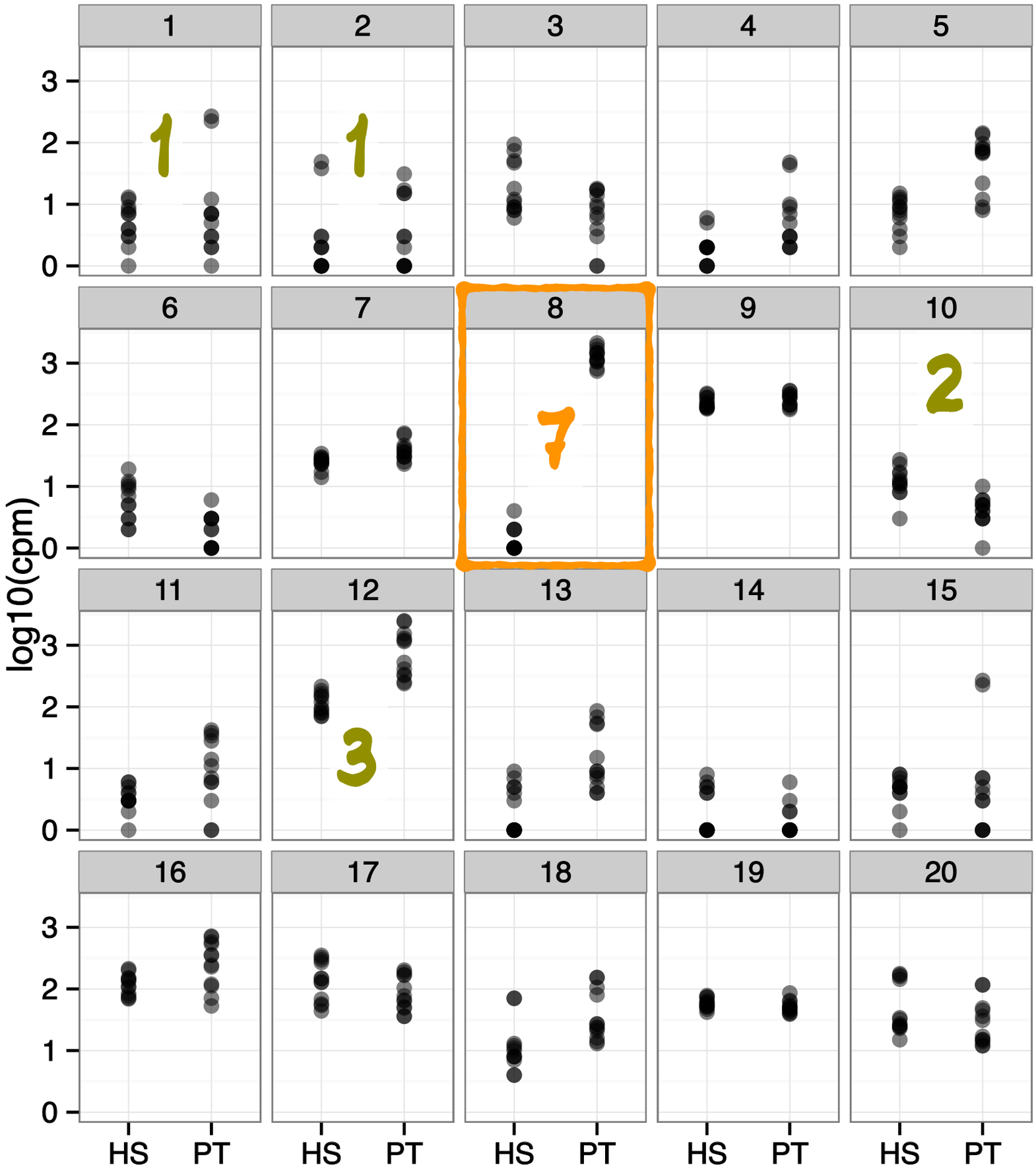
2nd

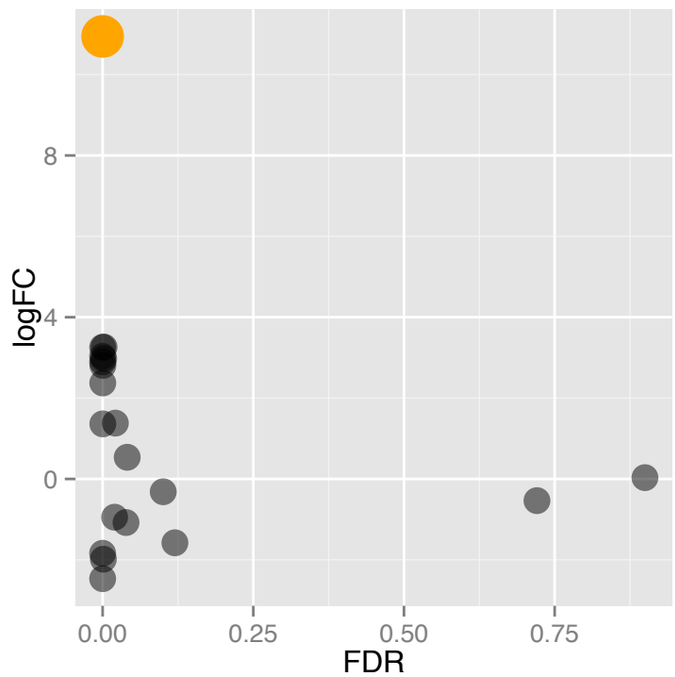




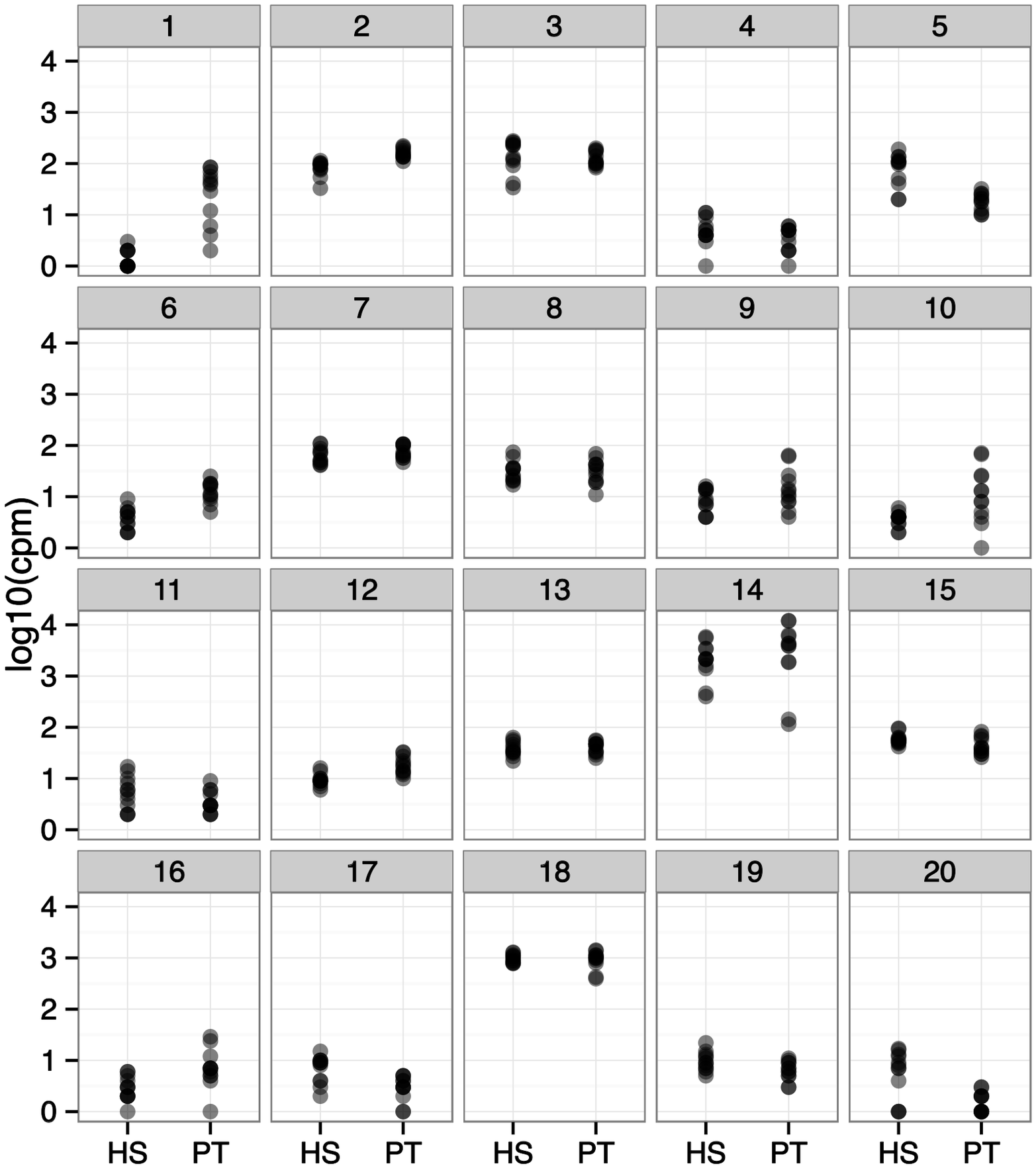
Human-chimp 1

2nd

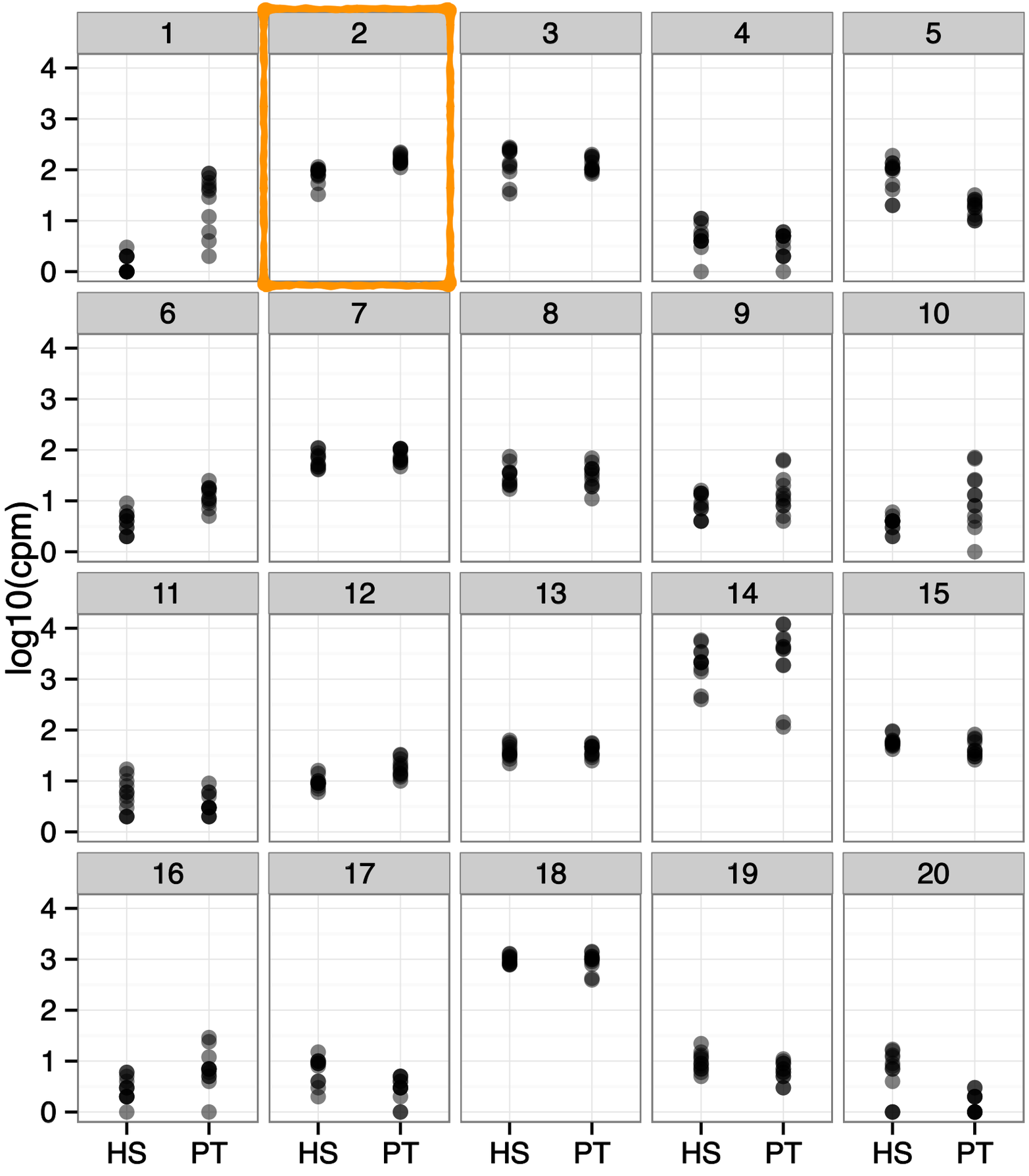


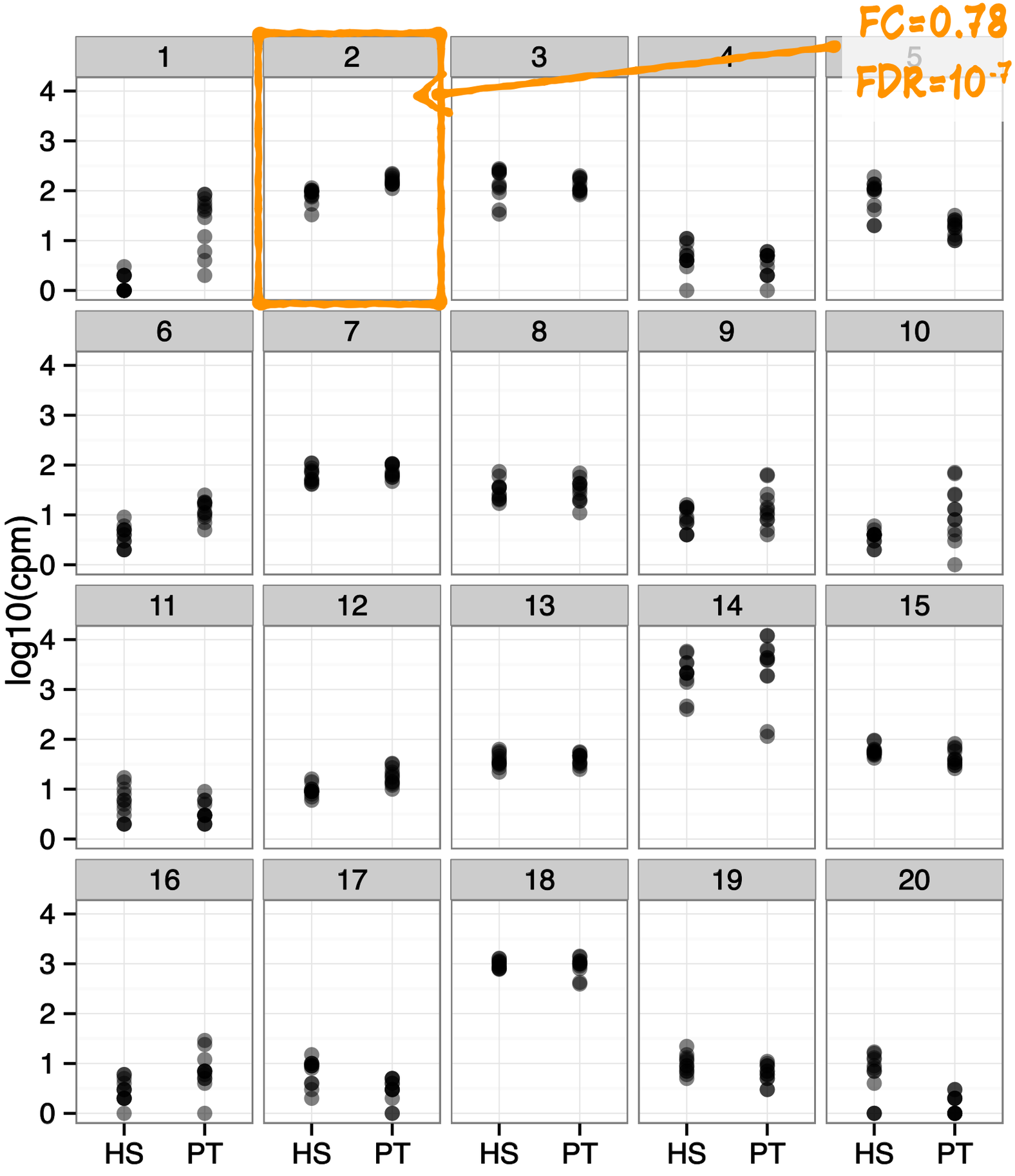


1000th

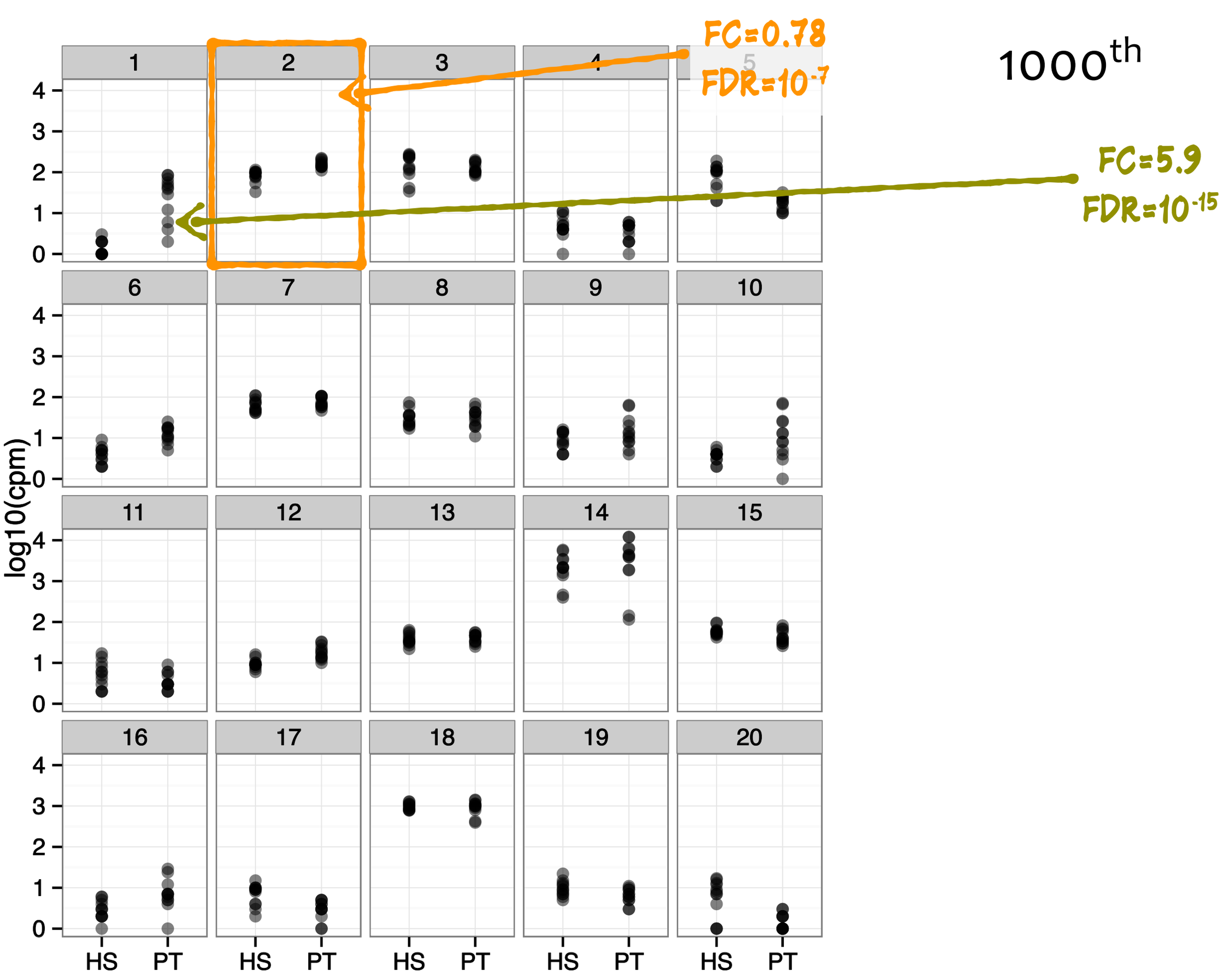


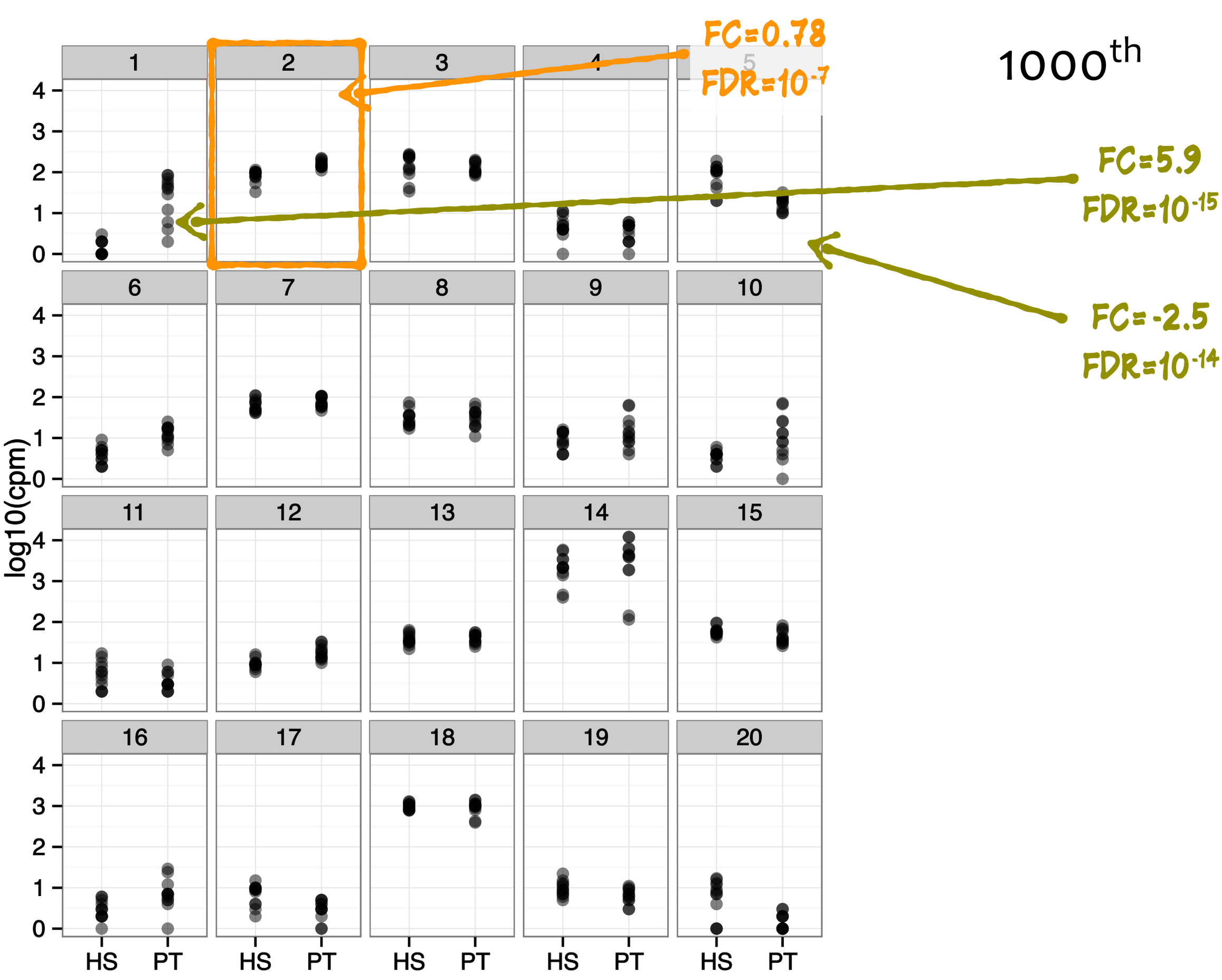
1000th




FC=0.78
FDR=10-7
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
log10(cpm)
0
1
2
3
4
HS
PT

1000th


FC=0.78
FDR=10-7
FC=5.9
FDR=10-15
1000th
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
log10(cpm)
0
1
2
3
4
HS
PT

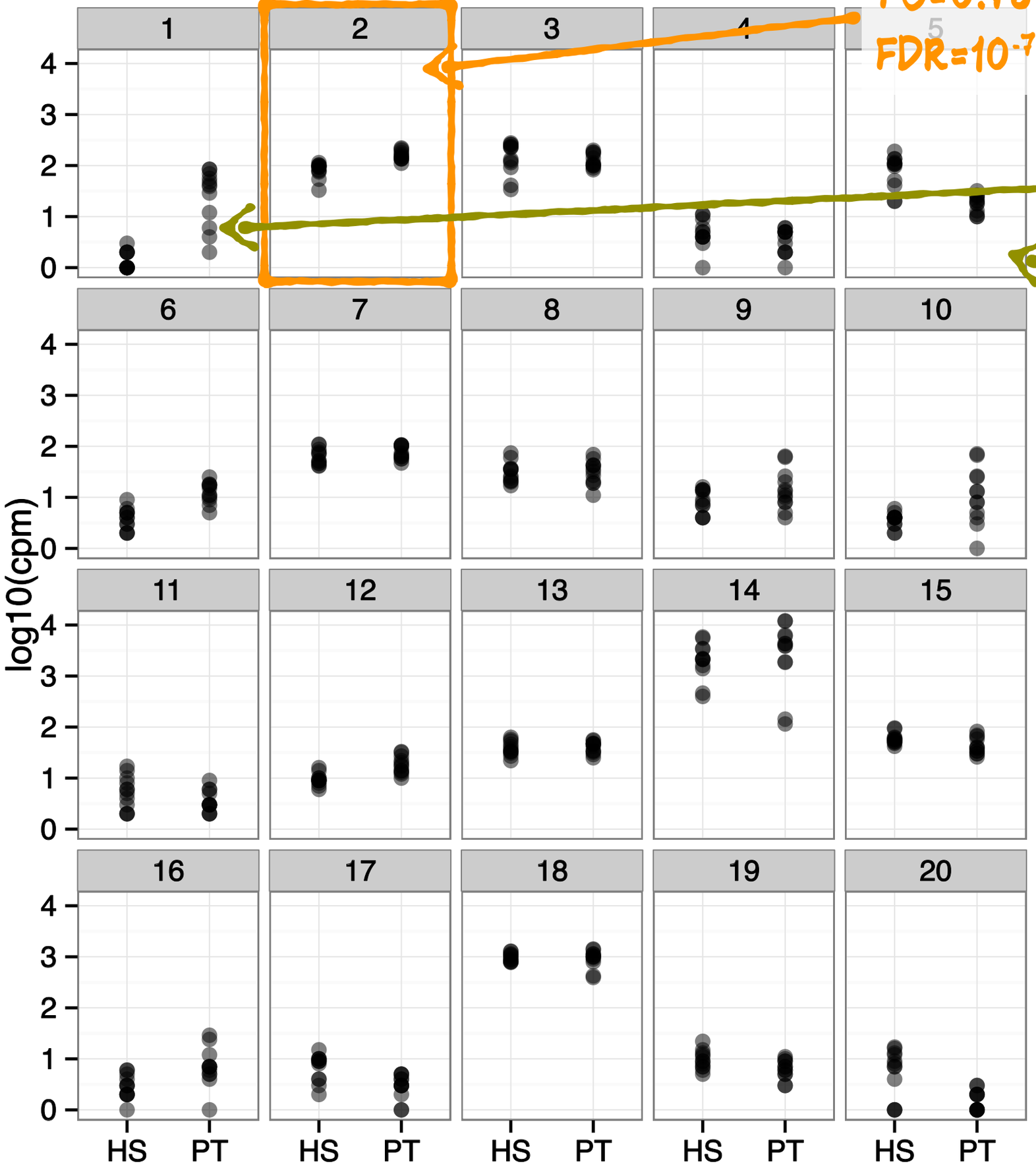

1000th
FC=0.78
FDR=10-7
FC=5.9
FDR=10-15
FC=-2.5
FDR=10-14
log10(cpm)
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
HS
PT

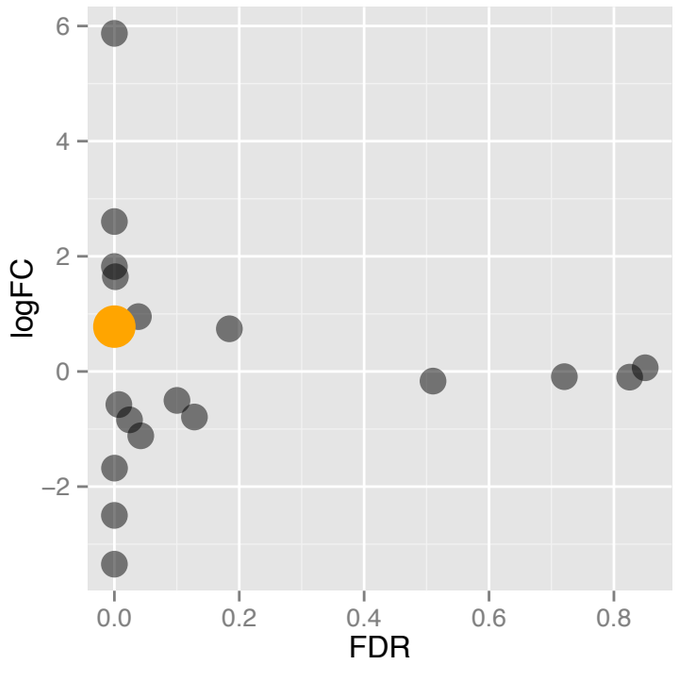
FC=0.78
FDR=10^{-7}

1000th

FC=5.9
FDR=10^{-15}

FC=-2.5
FDR=10^{-14}

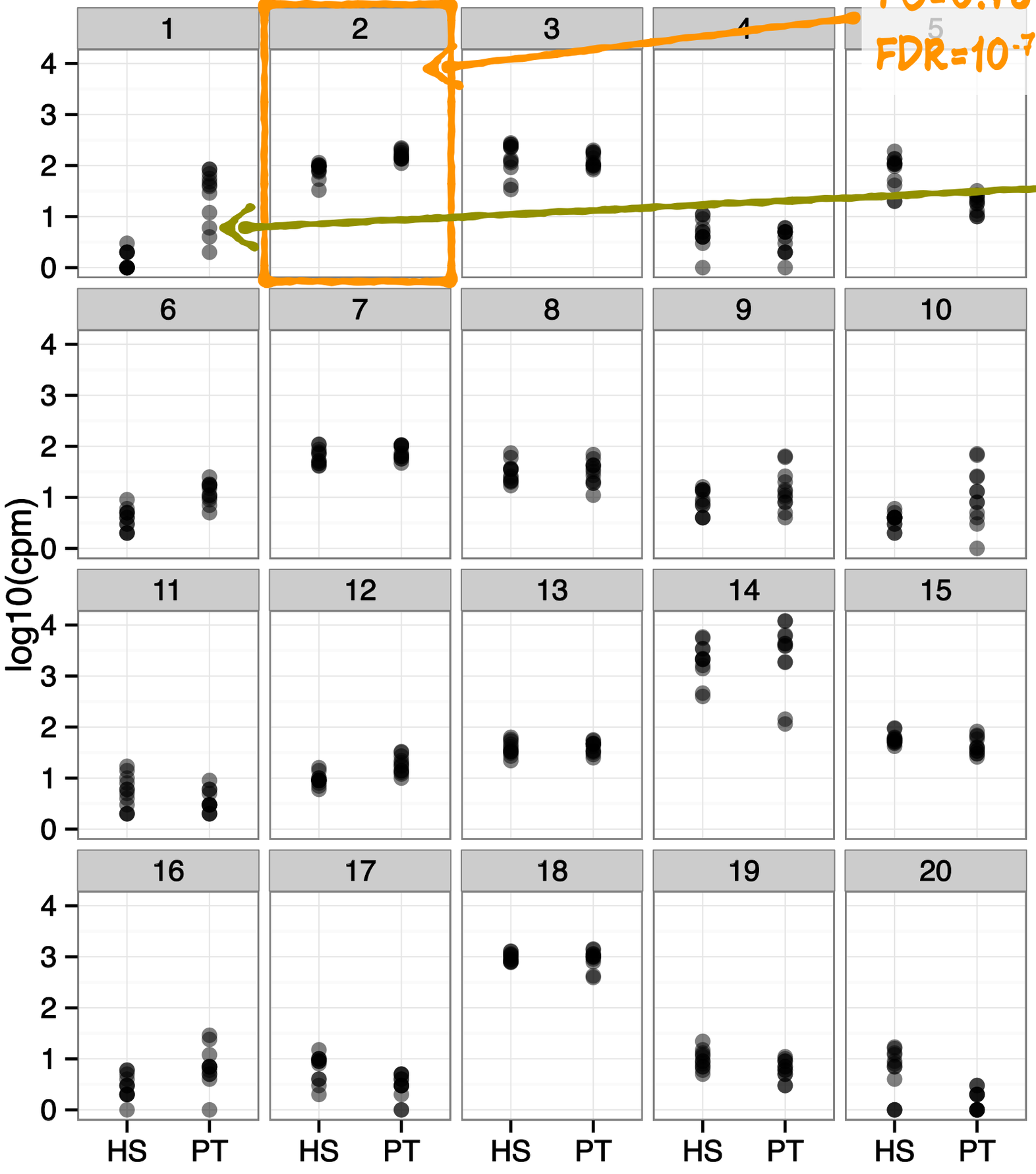


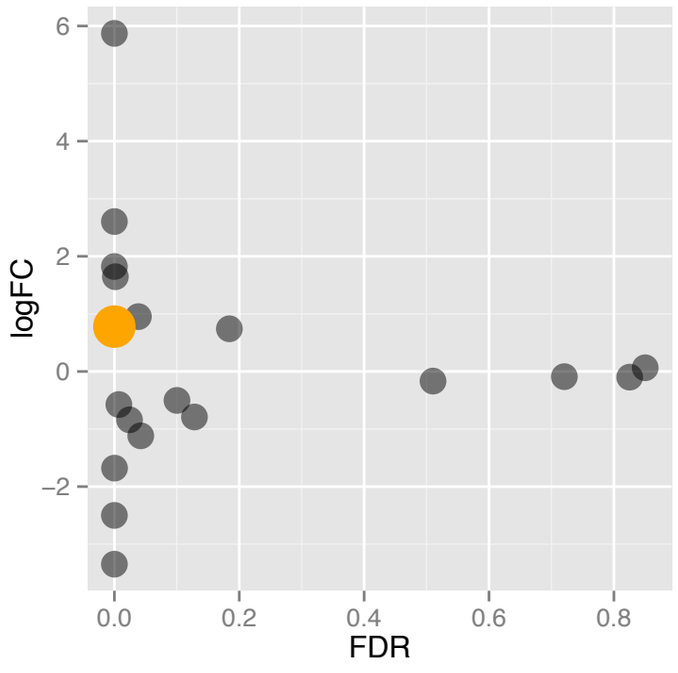


FC=0.78
FDR=10^{-7}

1000th

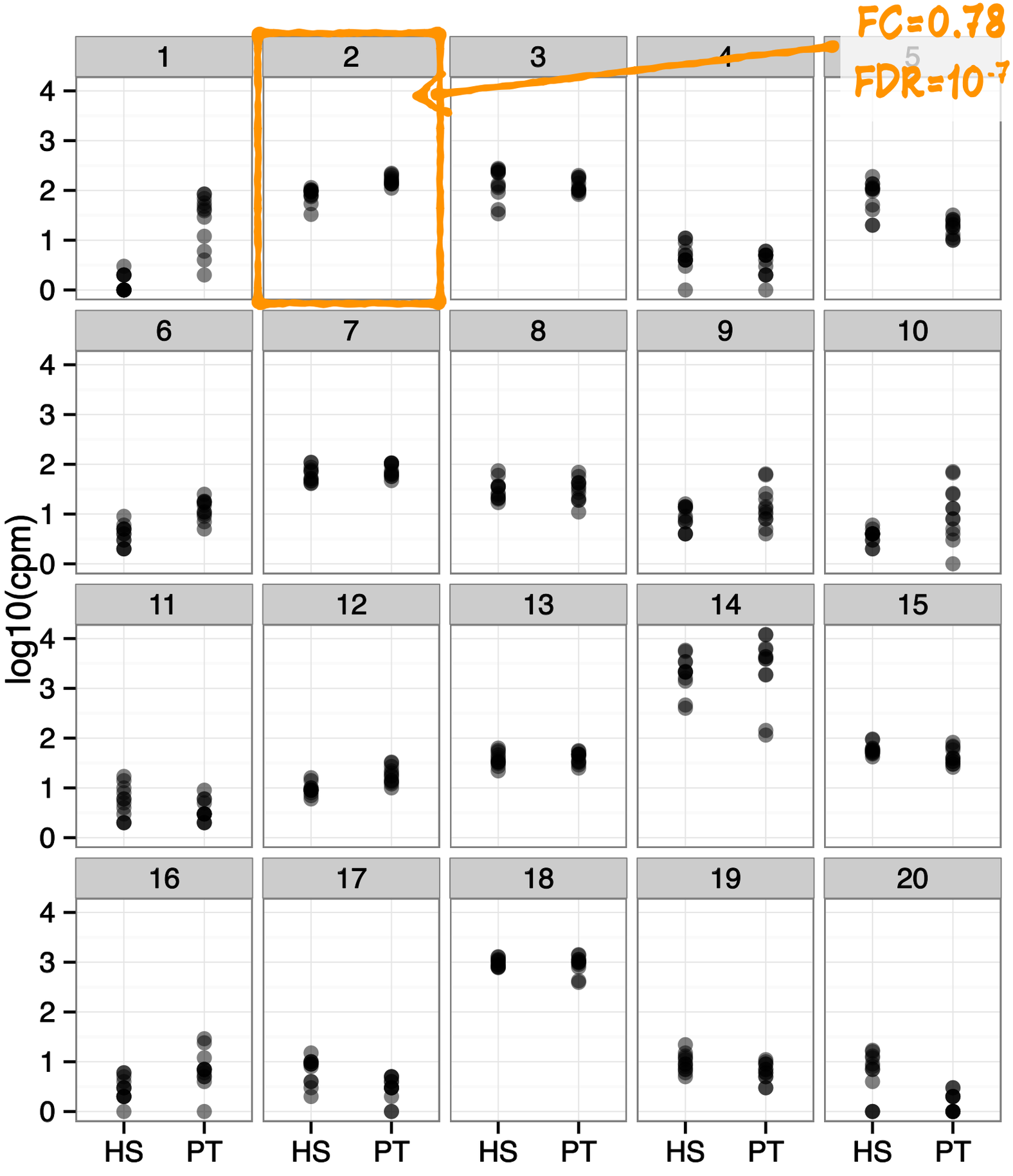
FC=5.9
FDR=10^{-15}

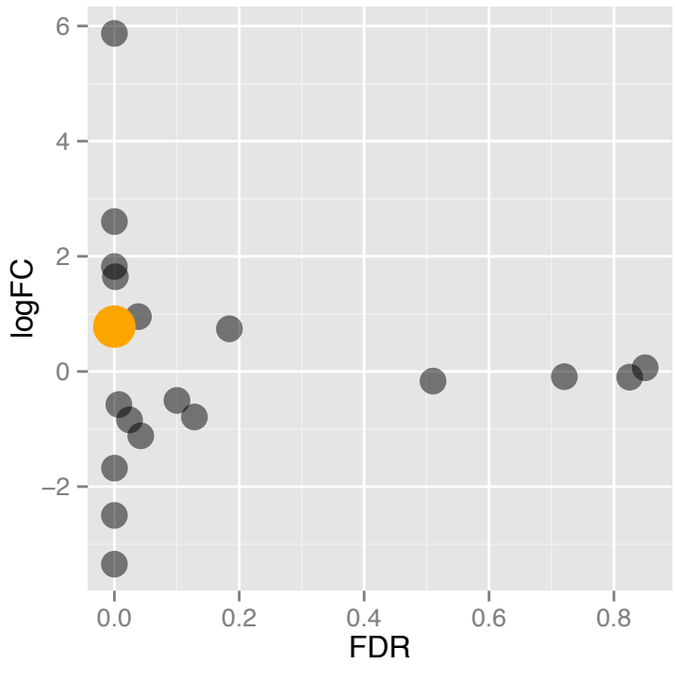




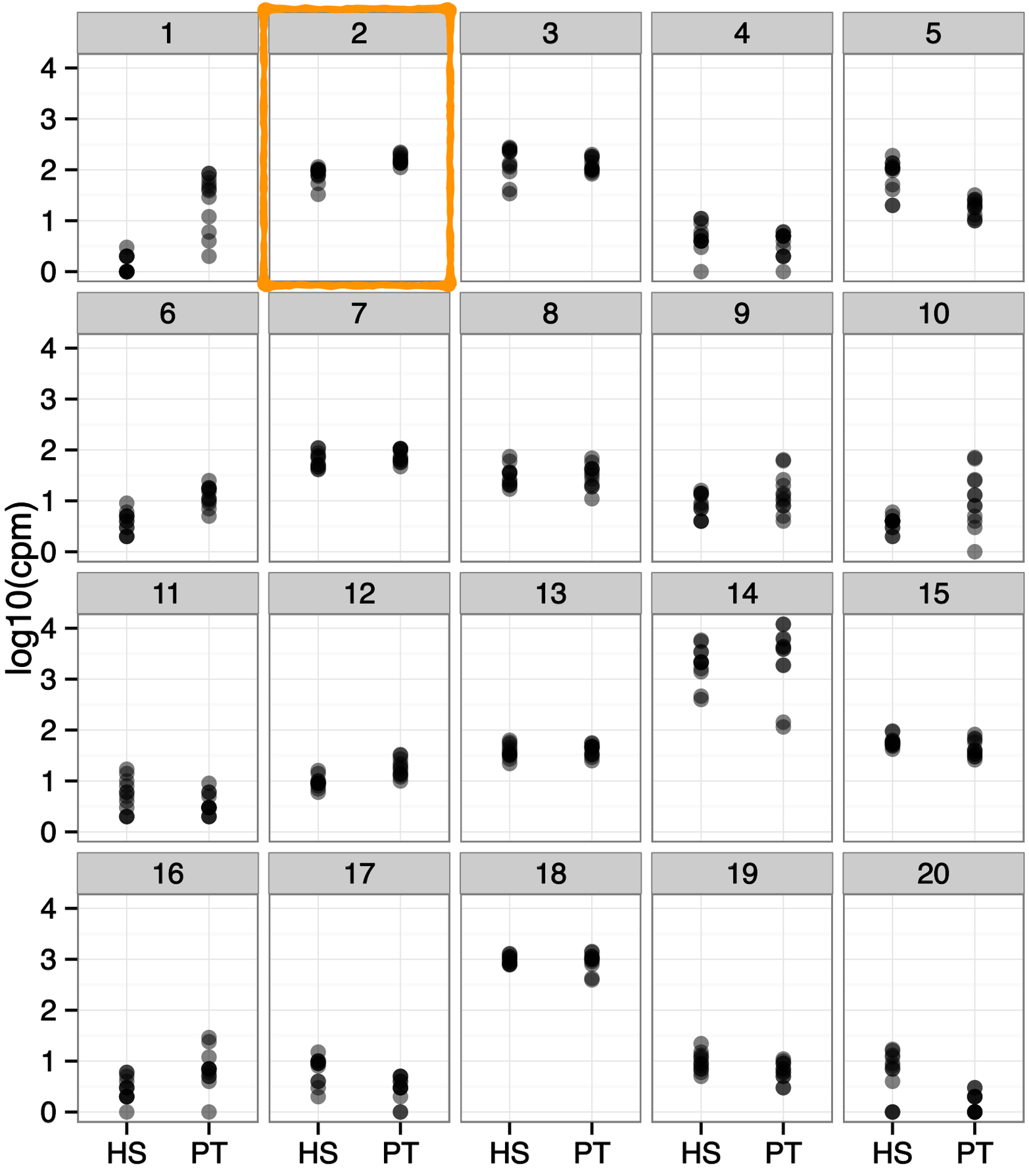
FC=0.78
FDR=10^{-7}

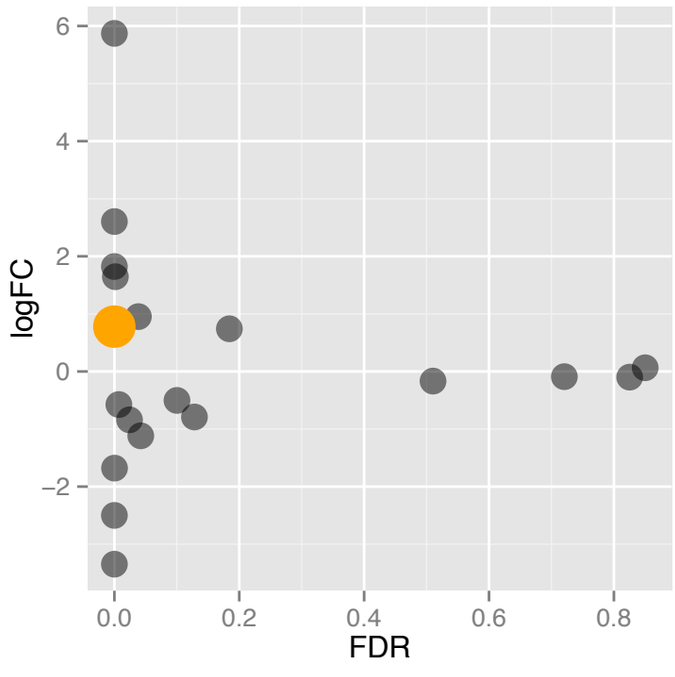
1000th



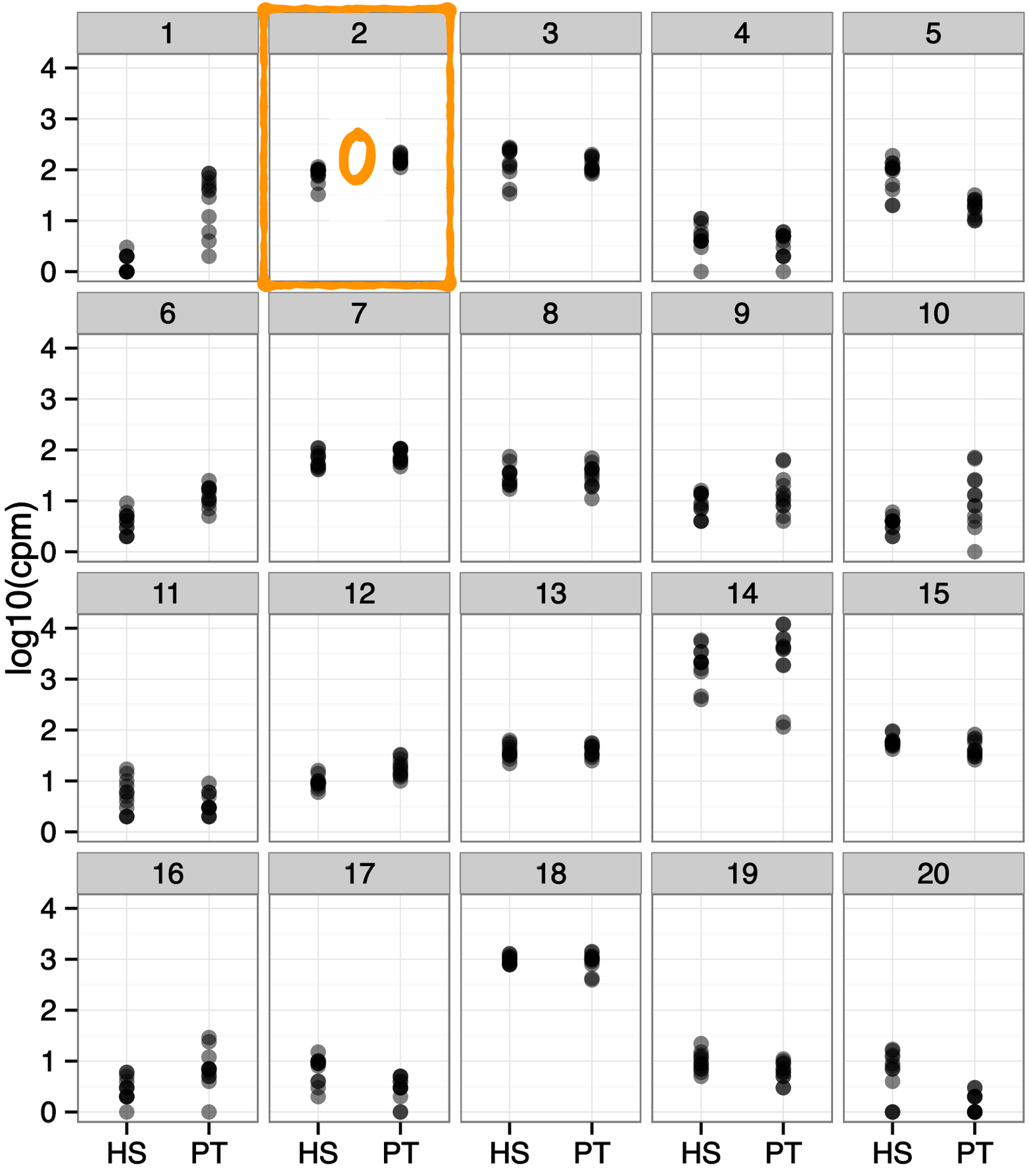


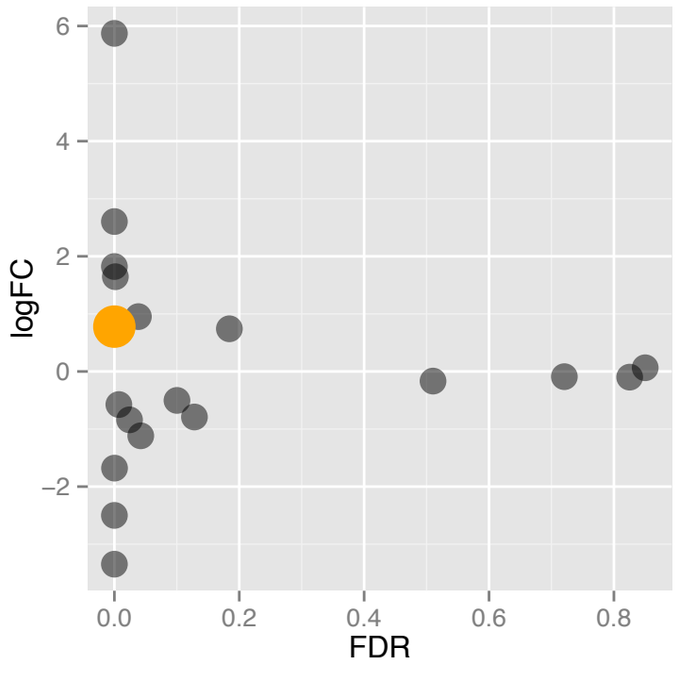
1000th


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
log10(cpm)
0
1
2
3
4
HS
PT

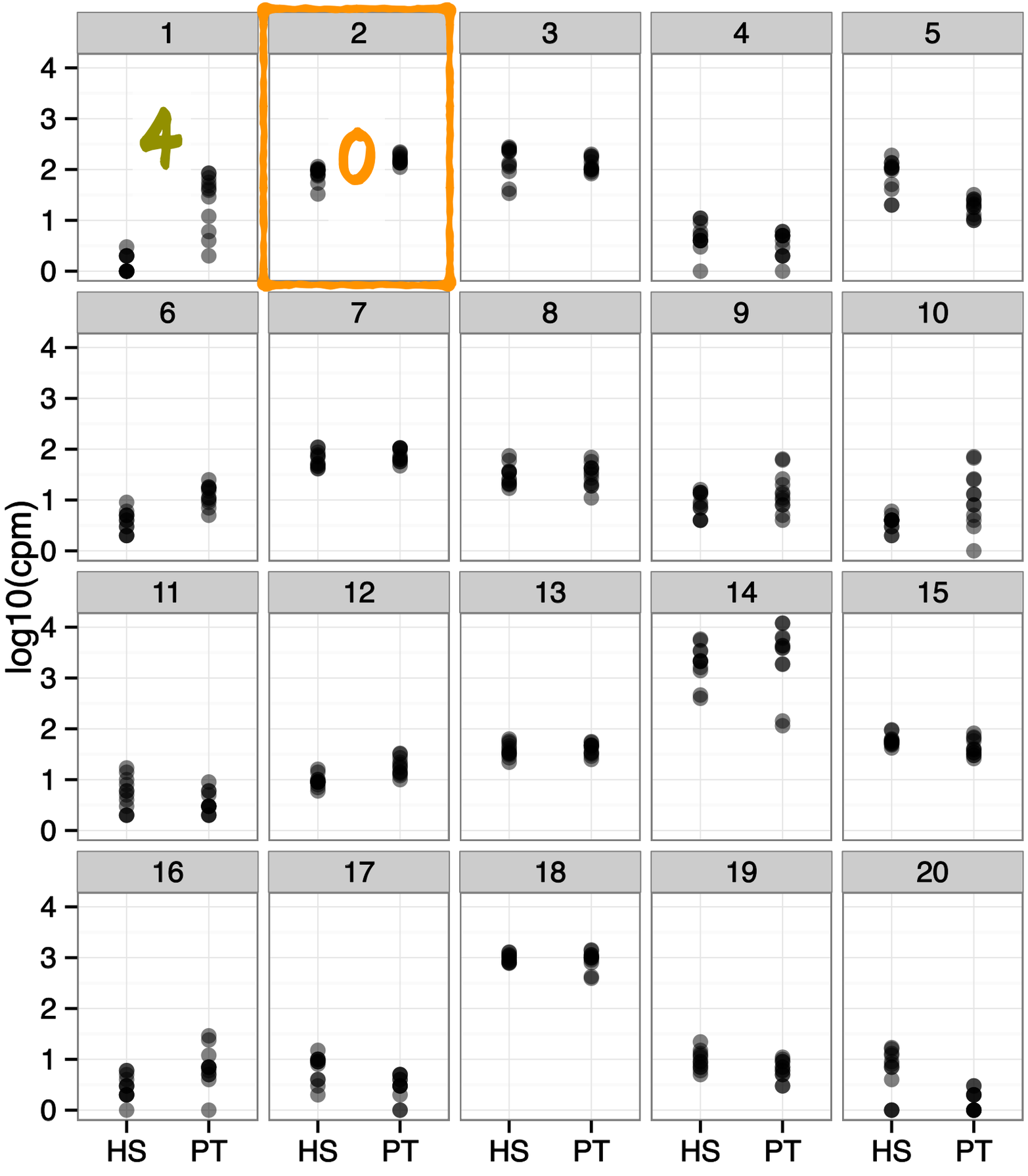

logFC
6
4
2
0
−2
0.0
0.2
0.4
0.6
0.8
FDR

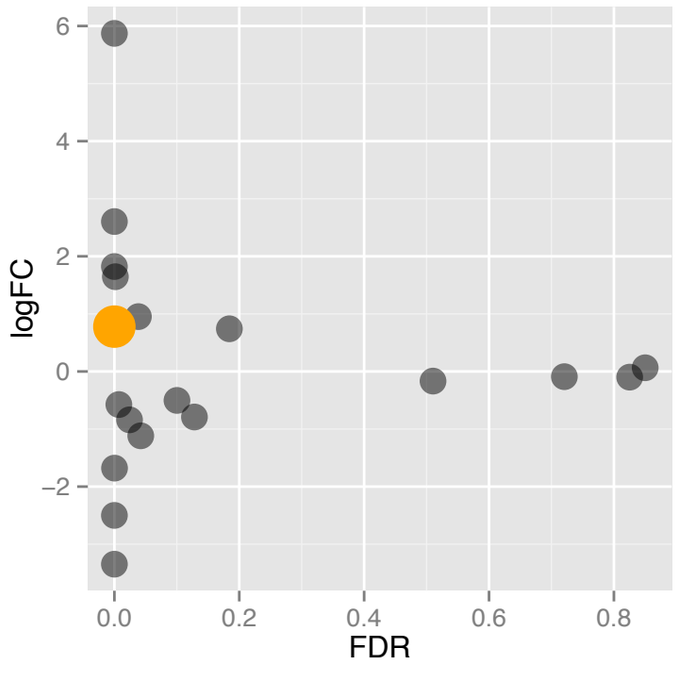
1000th


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
0
log10(cpm)
HS
PT

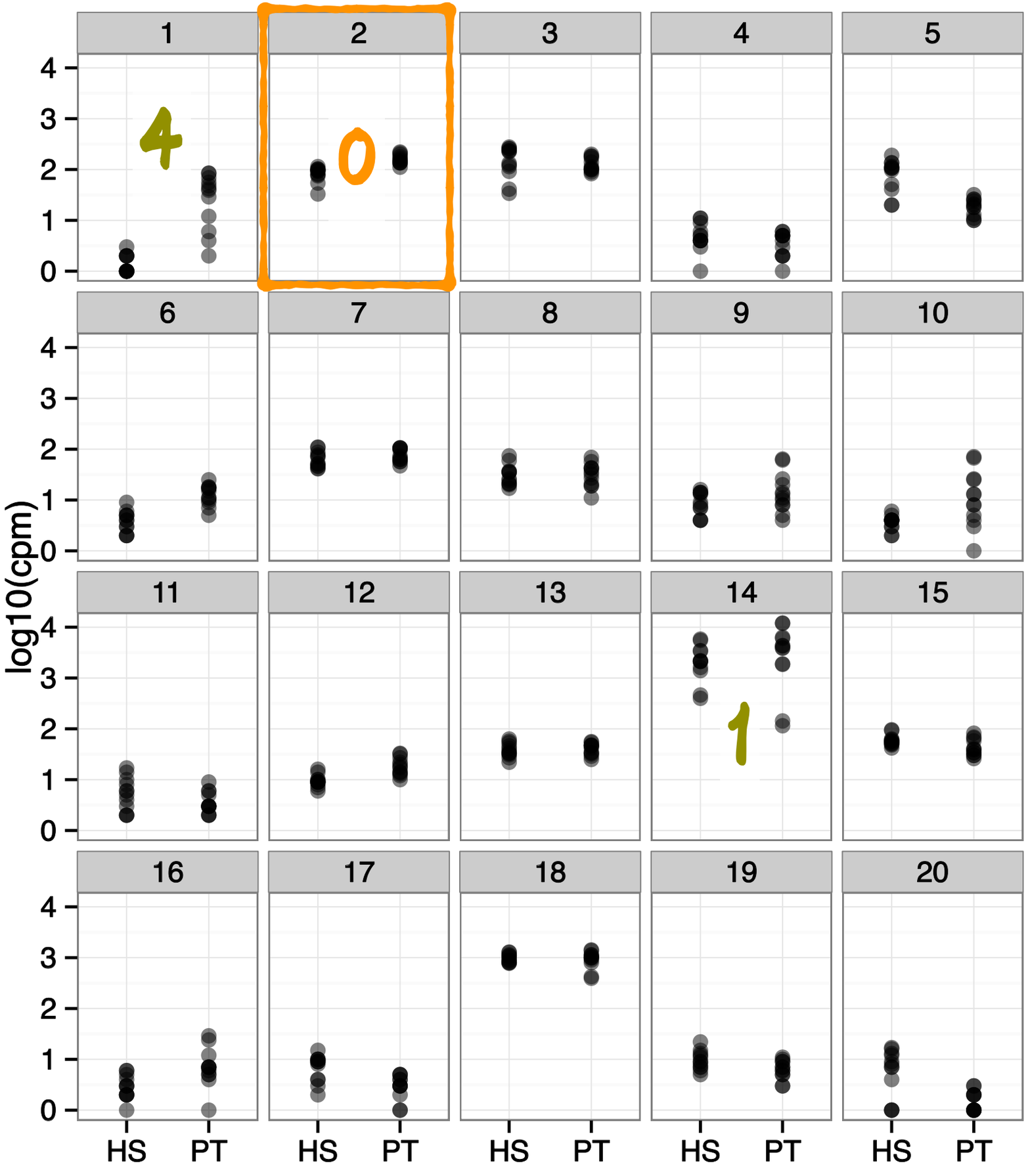

logFC
FDR
6
4
2
0
−2
0.0
0.2
0.4
0.6
0.8

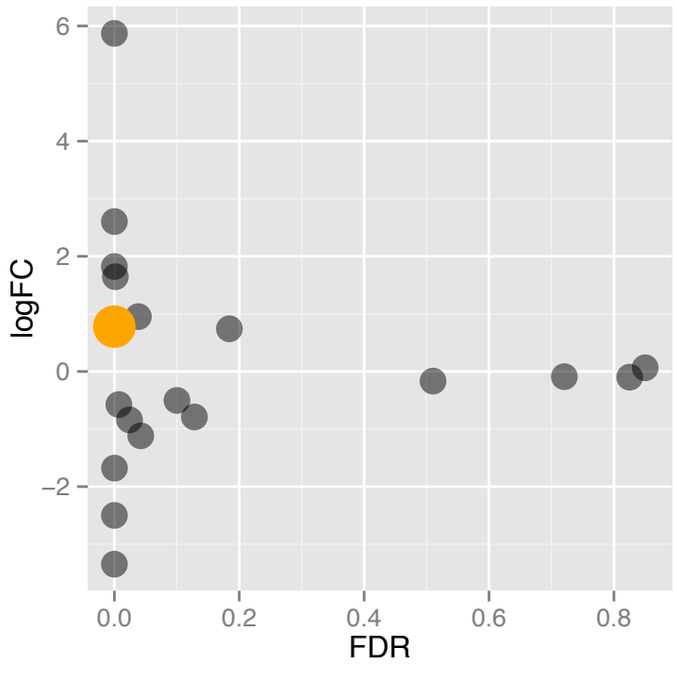
1000th


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
4
0
log10(cpm)
HS
PT


logFC
FDR
6
4
2
0
−2
0.0
0.2
0.4
0.6
0.8

1000th


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
log10(cpm)
HS
PT
4
0
1


logFC
FDR
6
4
2
0
−2
0.0
0.2
0.4
0.6
0.8

What we learn

- We see FDR-adjusted p-values as low as 10^{-14} in random assignment of treatment labels
- Would argue that the difference in expression maxes out around 100 genes

Costs

- 82 lineups
- 107 turkers
- 1078 lineups evaluated
- $115 total cost, $1/lineup, $10 Amazon fee, $5 for bonus’

Summary

- The lineup provides a rigorous way to evaluate data plots
- With the use of crowd-sourcing it could be used for high-throughput data
- Consideration of scale used, similar to tag-wise, trended or common
- Need to get some funding to pursue this!

Other topics: R packages

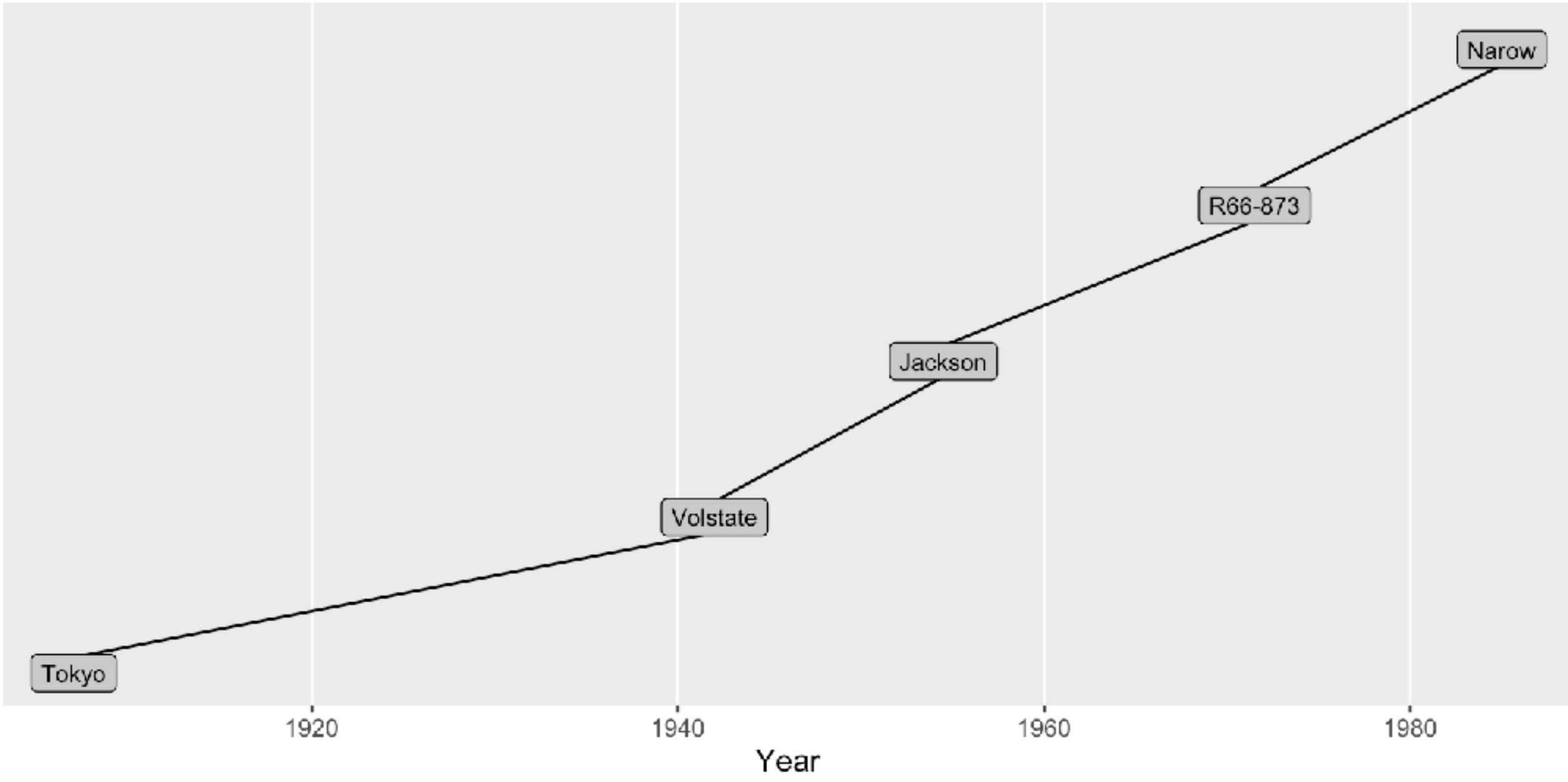
- genealogy: lineage relationships
- shiny apps for big data

ggenealogy

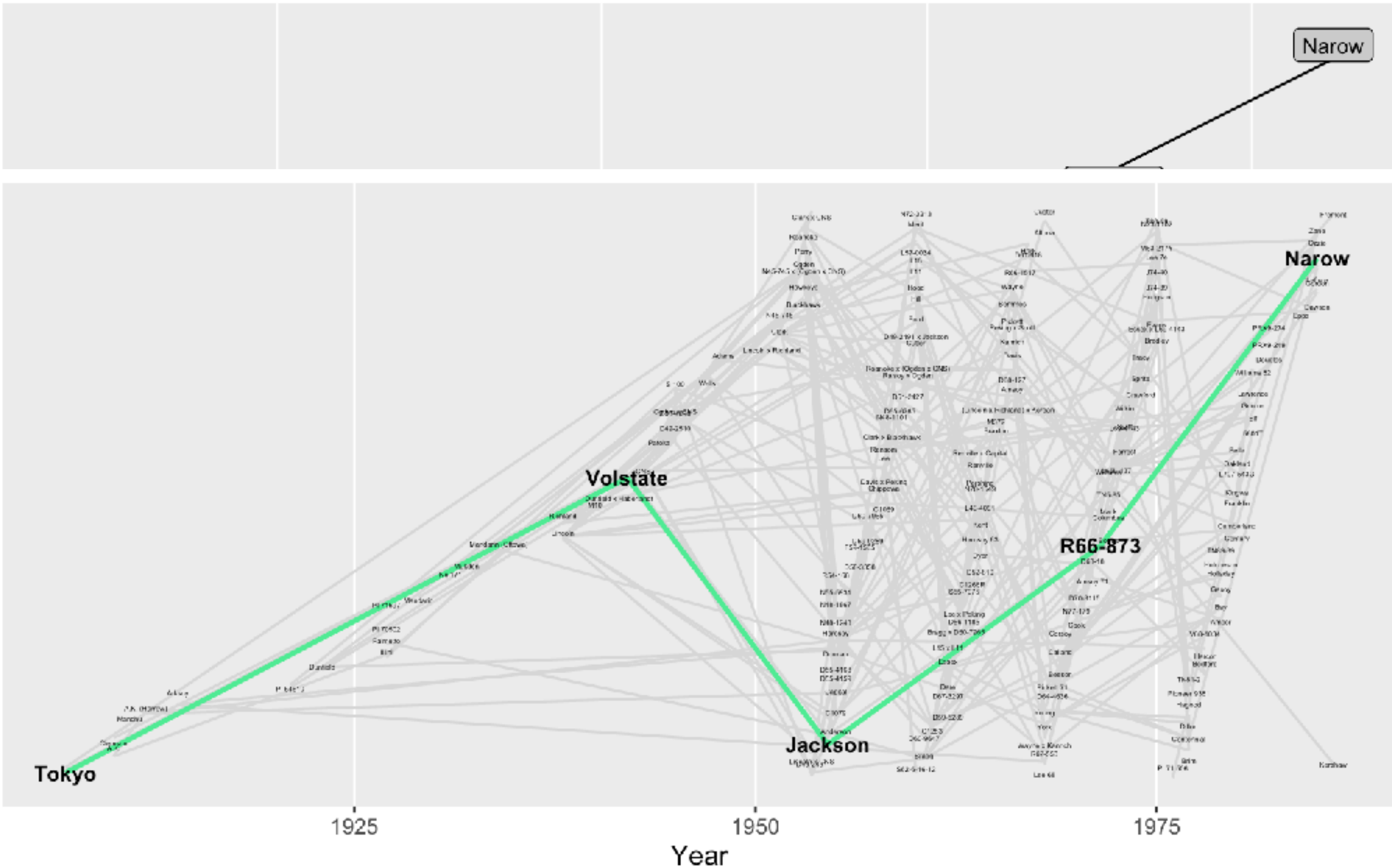
- shortest path
- plotting ancestors and descendants
- plotting distance matrix
- using interaction

```
pathTN <- getPath("Tokyo", "Narow", sbIG, sbGeneal)
pathTN
#> $pathVertices
#> [1] "Tokyo"    "Volstate" "Jackson"  "R66-873"  "Narow"
#>
#> $yearVertices
#> [1] "1907"   "1942"   "1954.5" "1971.5" "1985"
plotPath(pathTN)
```

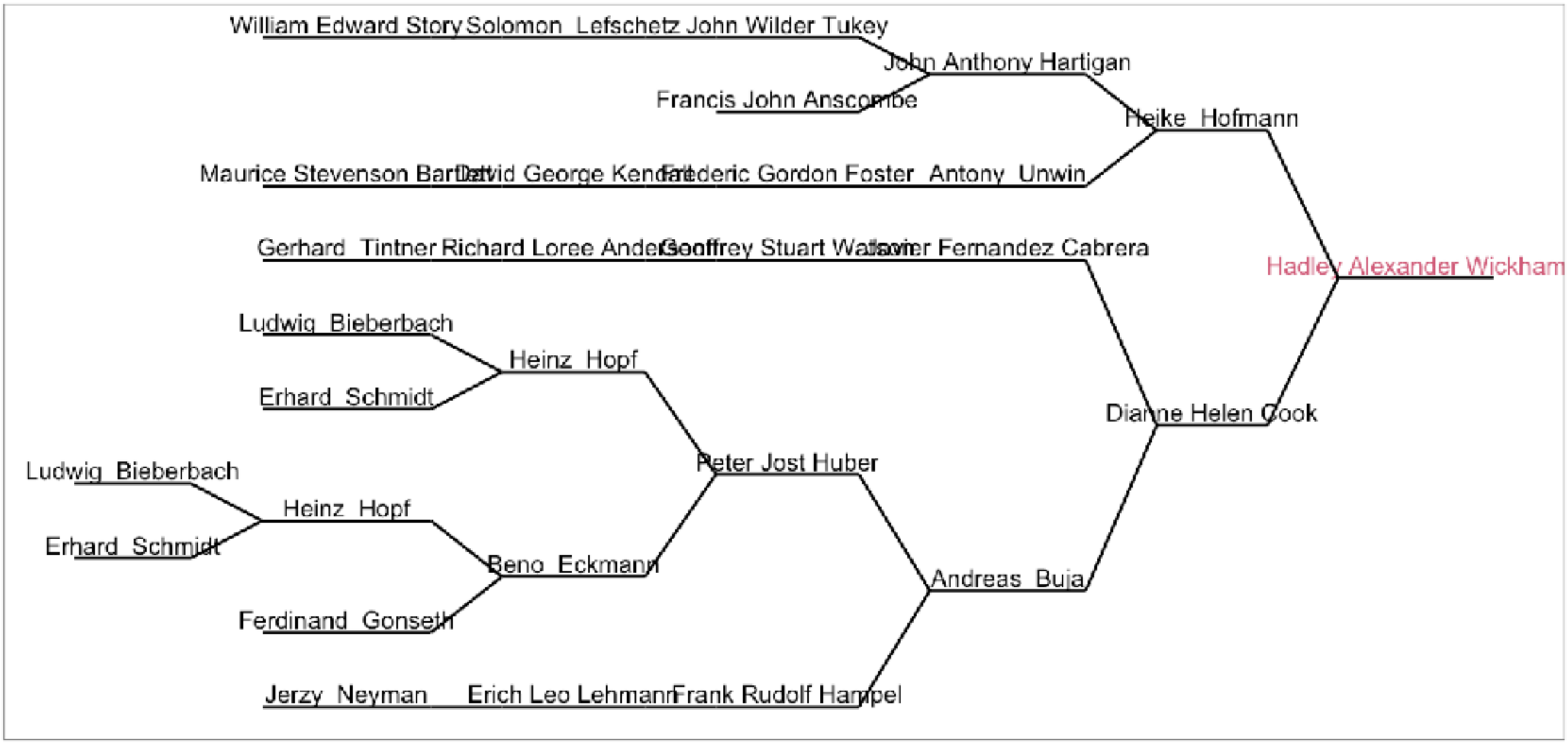
```
pathTN <- getPath("Tokyo", "Narow", sbIG, sbGeneal)
pathTN
#> $pathVertices
#> [1] "Tokyo"    "Volstate" "Jackson"  "R66-873"  "Narow"
```



```
pathTN <- getPath("Tokyo", "Narow", sbIG, sbGeneal)
pathTN
#> $pathVertices
#> [1] "Tokyo"    "Volstate" "Jackson"  "R66-873"  "Narow"
```



```
hw <- read_csv("../data/hw-gen.csv")
names(hw)[2:3] <- c("parent", "child")
plotAncDes("Hadley Alexander Wickham", hw, mAnc=6, mDes=1)
```



For more information go to:

https://github.com/dicook/SISBID-2016

Shiny apps for big data

- Explore genetic signatures, genealogy and phenotypic changes of soybean breeding
- Understand how the genome changed with the breeding of lines, and how this affected other traits
- Data sources:
  - Next-generation sequencing DNA-seq on 79 lines: DNA sequencing libraries were prepared using TruSeq DNA sample prep and NuGENs unamplified prep kits (Illumina Inc., San Diego, CA and NuGEN Technologies Inc., San Carlos, CA).
  - Field yield trials: 30/79 + 138 ancestral lines
  - Breeding literature, what lines were bred to produce what line

- Copy number variation (CNV): 2Gb of analysis files, annotations
  - Seven tabs containing different functionality
  - Four of the tabs, CNV Location, Copy Number, "Search CNVs by Location", and CNV List, primarily concerned with exploring the identified copy number variants
  - The other three tabs, Phenotype Data, Genealogy, and Methodology provide additional information about the soybean cultivars and the experimental methodology
- SNPs: 12Gb of data, 20mill SNPs, 1mil locations, 79 lines
- Genealogy: Shows the parent to child lineage

Apps written by Dr Susan Vanderplas

- Copy number variation (CNV): 2Gb of analysis files, annotations `http://shiny.soybase.org/CNV/`
  - Seven tabs containing different functionality
  - Four of the tabs, CNV Location, Copy Number, "Search CNVs by Location", and CNV List, primarily concerned with exploring the identified copy number variants
  - The other three tabs, Phenotype Data, Genealogy, and Methodology provide additional information about the soybean cultivars and the experimental methodology
- SNPs: 12Gb of data, 20mill SNPs, 1mil locations, 79 lines
- Genealogy: Shows the parent to child lineage

Apps written by Dr Susan Vanderplas

Inference References

- Buja et al (2009) R. Soc. Phil. Trans. A
- Hofmann et al (2012) IEEE TVCG
- Majumder et al (2013) JASA
- Yin et al (2013) J. Data Mining in Gen. & Prot.
- Roy Chowdhury et al (2014) Computational Statistics