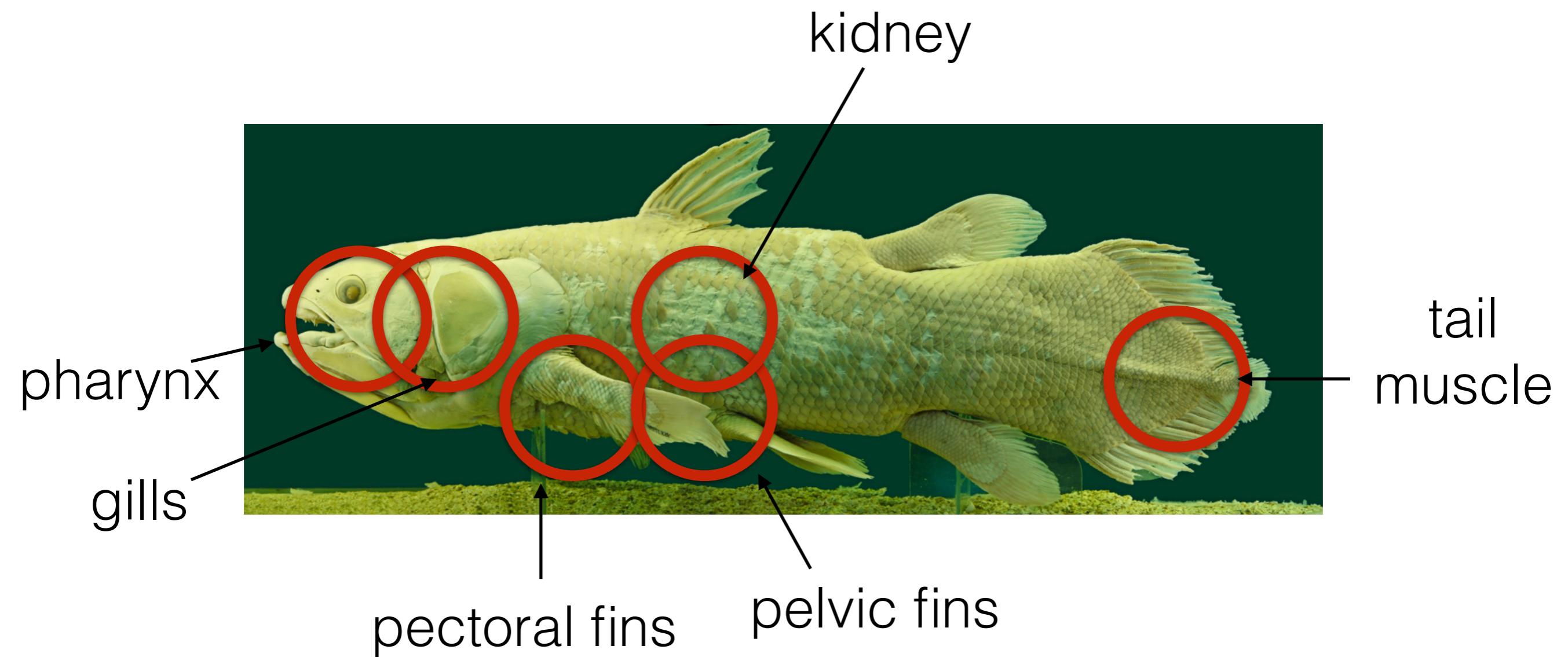


Live sleuth demo

Harold Pimentel

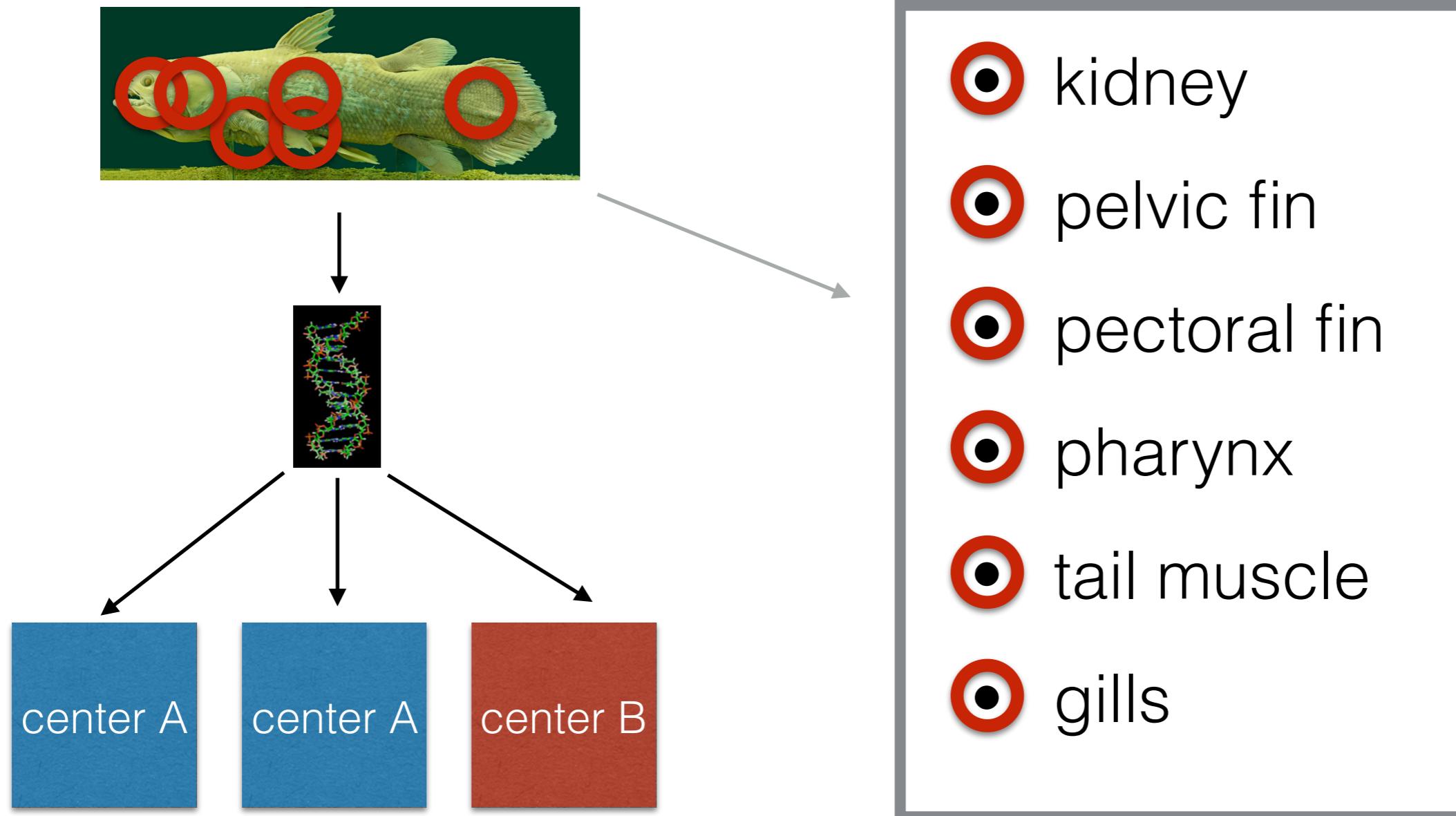
Coelacanth analysis



Sample to condition table

	sample	center	tissue	path
1	DRR002302	NIG	kidney	results/DRR002302/kallisto
2	DRR002303	NIG	pelvic fin	results/DRR002303/kallisto
3	DRR002304	NIG	pectoral fin	results/DRR002304/kallisto
4	DRR002305	NIG	pharynx	results/DRR002305/kallisto
5	DRR002306	NIG	gill	results/DRR002306/kallisto
6	DRR002307	NIG	tail muscle	results/DRR002307/kallisto
7	DRR002308	NIG	kidney	results/DRR002308/kallisto
8	DRR002309	NIG	pelvic fin	results/DRR002309/kallisto
9	DRR002310	NIG	pectoral fin	results/DRR002310/kallisto
10	DRR002311	NIG	pharynx	results/DRR002311/kallisto
11	DRR002312	NIG	gill	results/DRR002312/kallisto
12	DRR002313	NIG	tail muscle	results/DRR002313/kallisto
13	DRR002314	NIG	tail muscle	results/DRR002314/kallisto
14	DRR002319	UT-MGS	gill	results/DRR002319/kallisto
15	DRR002320	UT-MGS	tail muscle	results/DRR002320/kallisto
16	DRR002317	UT-MGS	pectoral fin	results/DRR002317/kallisto
17	DRR002316	UT-MGS	pelvic fin	results/DRR002316/kallisto
18	DRR002315	UT-MGS	kidney	results/DRR002315/kallisto
19	DRR002318	UT-MGS	pharynx	results/DRR002318/kallisto

Experimental design



Testing for tissue specific expression

- Ideally, we sample from **multiple** individuals

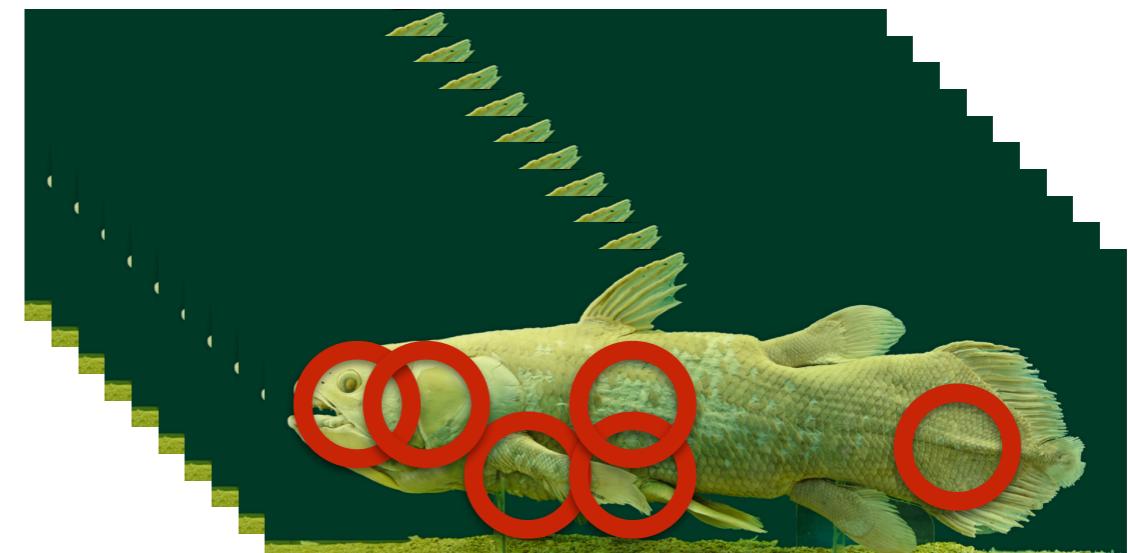
what we have

$N = 1$



what we want

$N = \text{many}$



Testing for tissue specific

- Ideally, we sample from **multiple** individuals

what we have

$N = 1$



what we settle for

$N = 3$

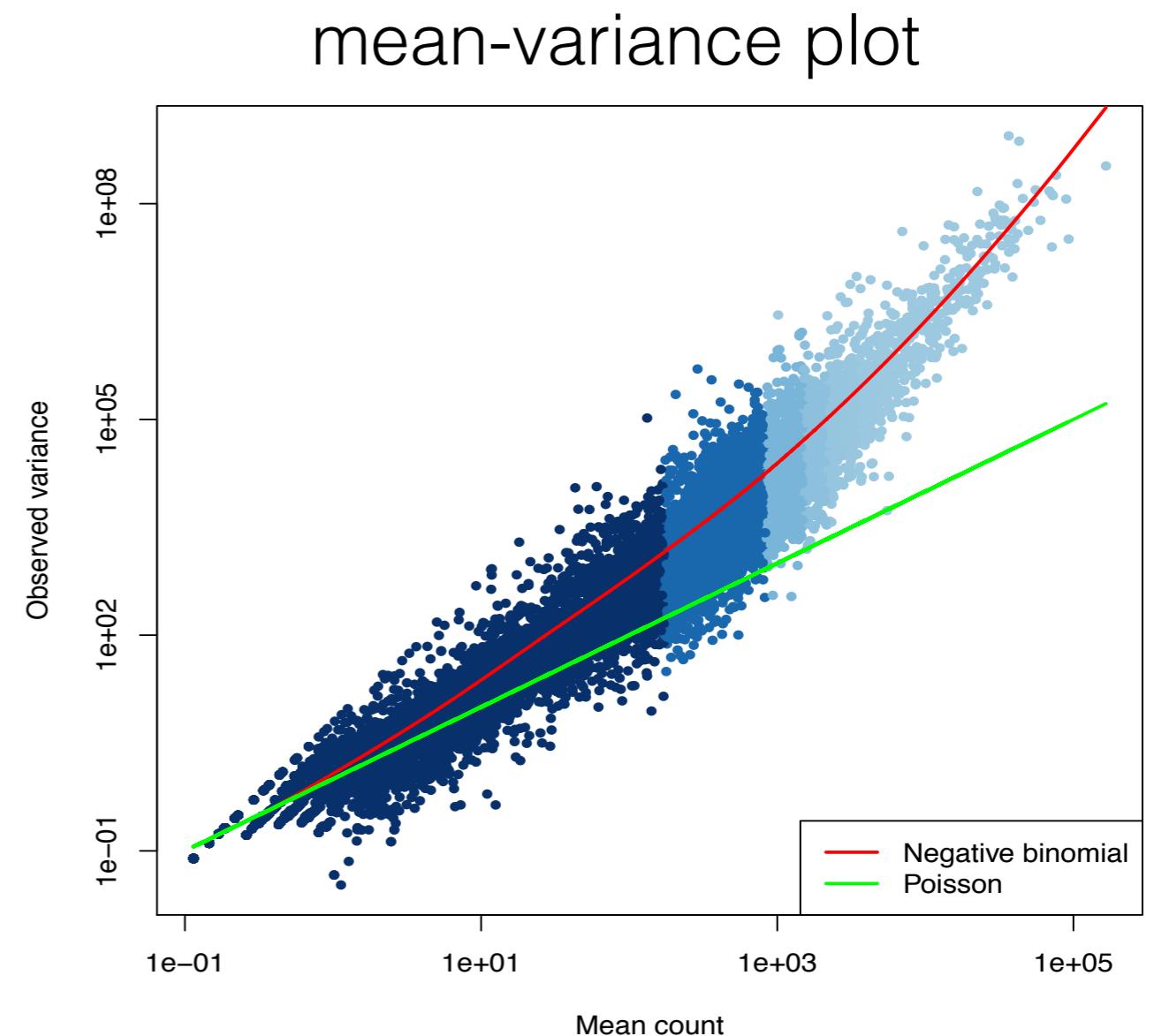


How to estimate variance with few replicates

- (Simple case) samples from two experimental conditions with few replicates
 - Which features likely have a different population mean?
- Essentially the classic *two-sample* problem with some complications:
 - **Transcript abundances must be *estimated* from transcript compatibility counts**
 - **Few replicates make variance estimation challenging**
 - **Multiple tests reduce power**

Shrinkage

- Make parametric assumption for distribution of counts
- Assume a smooth mean-variance count relationship to share information between transcripts (genes)
- Most methods do not incorporate "inferential variance" estimates



Sources of variation

- Technical variance
 - Library preparation
- Biological variance
 - Variation between individuals
- Inferential variance
 - Variation due to algorithms

Sources of variation

- Technical variance
 - Library preparation
- Biological variance
 - Variation between individuals
- Inferential variance
 - Variation due to algorithms

Sources of variation

- Technical variance
 - Library preparation
 - Biological variance
 - Variation between individuals
 - Inferential variance
 - Variation due to algorithms
- assumed to be fixed

Sources of variation

- Technical variance
 - Library preparation
- Biological variance
 - Variation between individuals
- Inferential variance
 - Variation due to algorithms

Sources of variation

- Technical variance
 - Library preparation
 - Biological variance
 - Variation between individuals
 - Inferential variance
 - Variation due to algorithms
- assumed to be fixed
- shrinkage estimation

Sources of variation

- Technical variance
 - Library preparation

assumed to be fixed

- Biological variance
 - Variation between individuals

shrinkage estimation

- Inferential variance
 - Variation due to algorithms

Sources of variation

- Technical variance
 - Library preparation

assumed to be fixed

- Biological variance
 - Variation between individuals

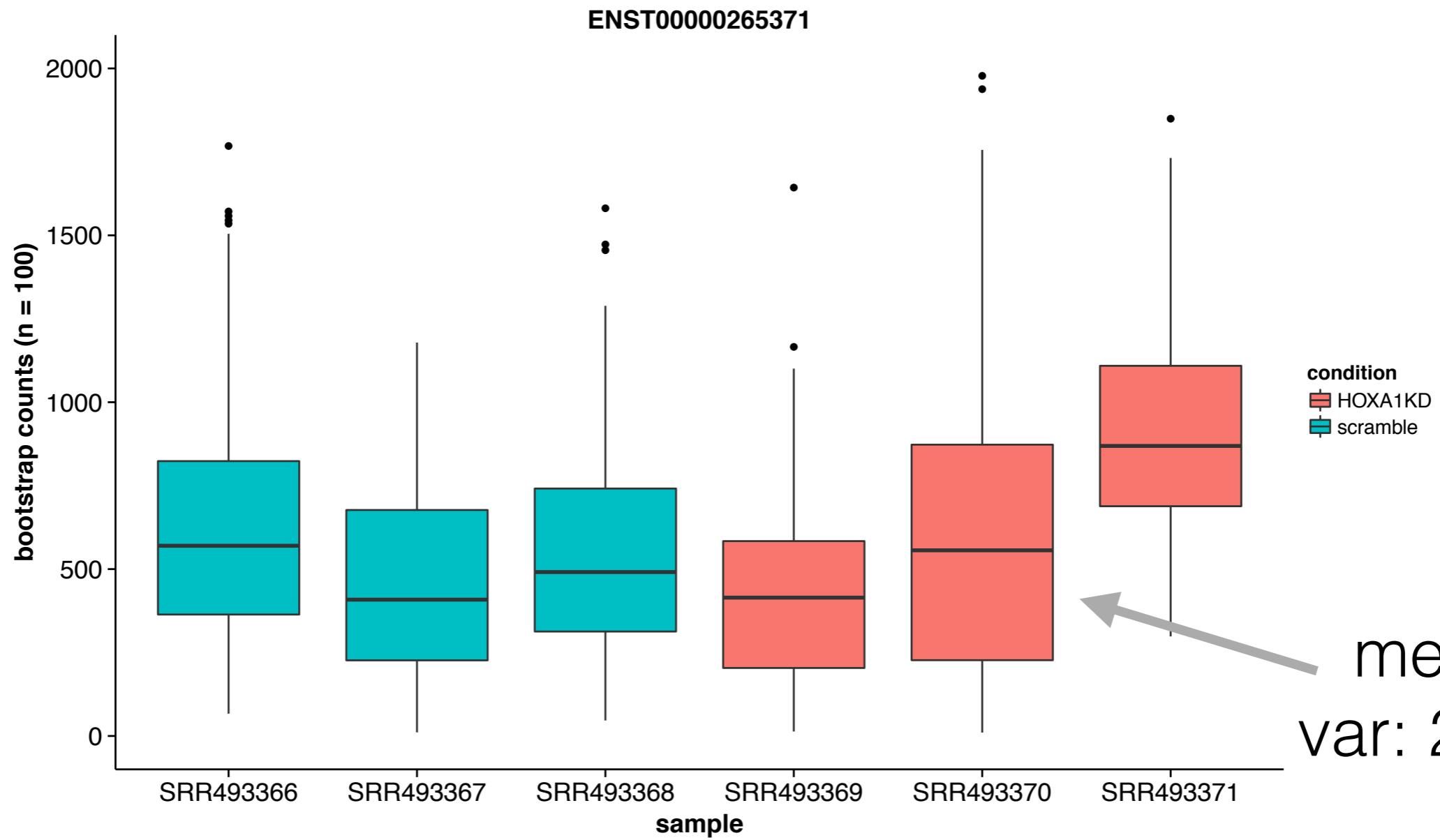
shrinkage estimation

- Inferential variance
 - Variation due to algorithms

normally ignored,
but modeled in sleuth

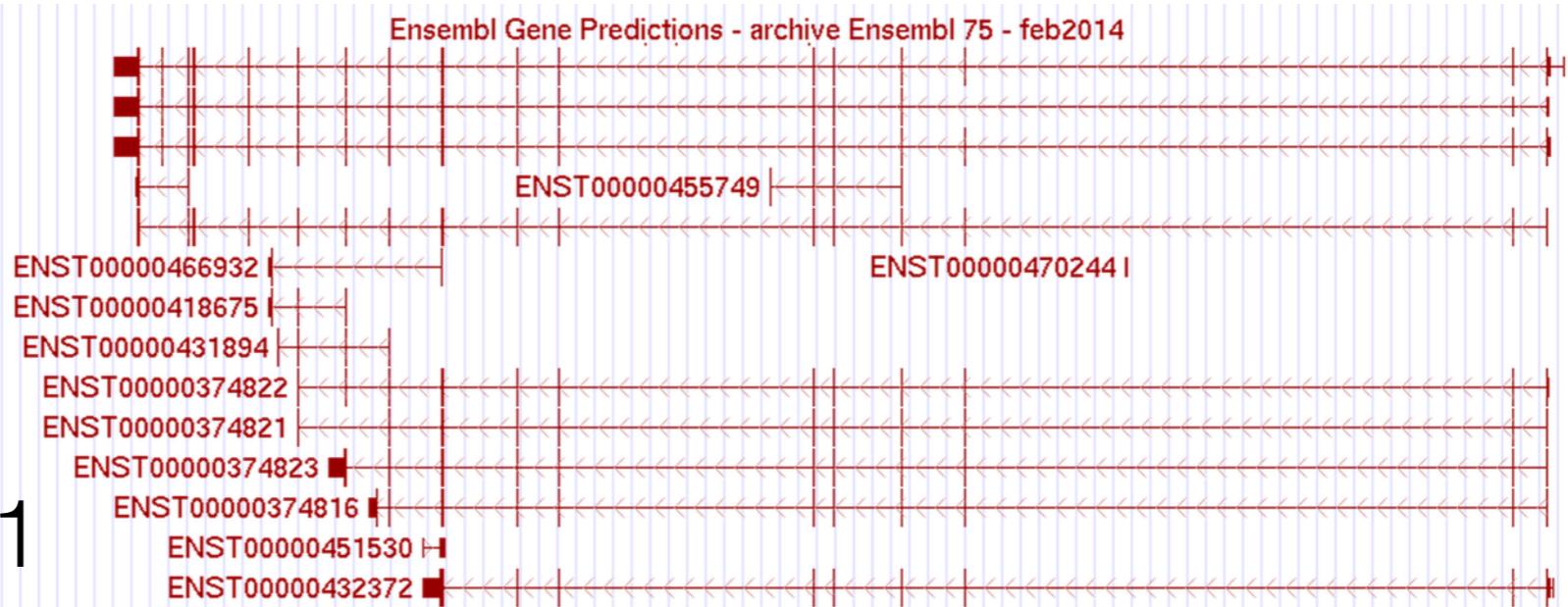
How much inferential variance?

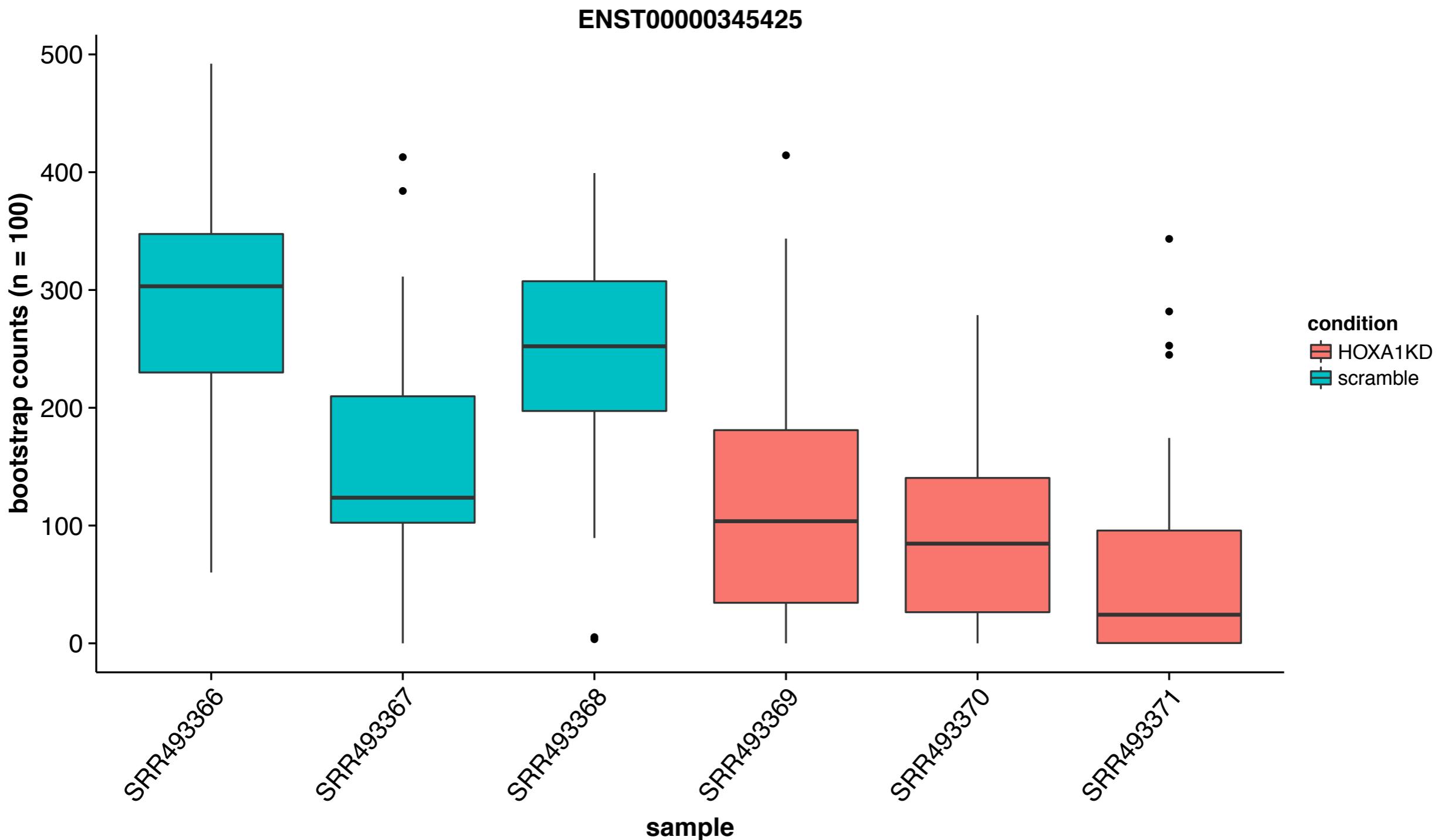
- Often, it is not so bad, but some genes are really complicated and short reads aren't that informative
- Case study:
 - Cuffdiff2 HiSeq data set ~15-20M fragments/sample
 - 3 HOXA1 knockdown in lung fibroblasts
 - 3 with scrambled siRNA
 - Ran *kallisto* with 100 bootstraps



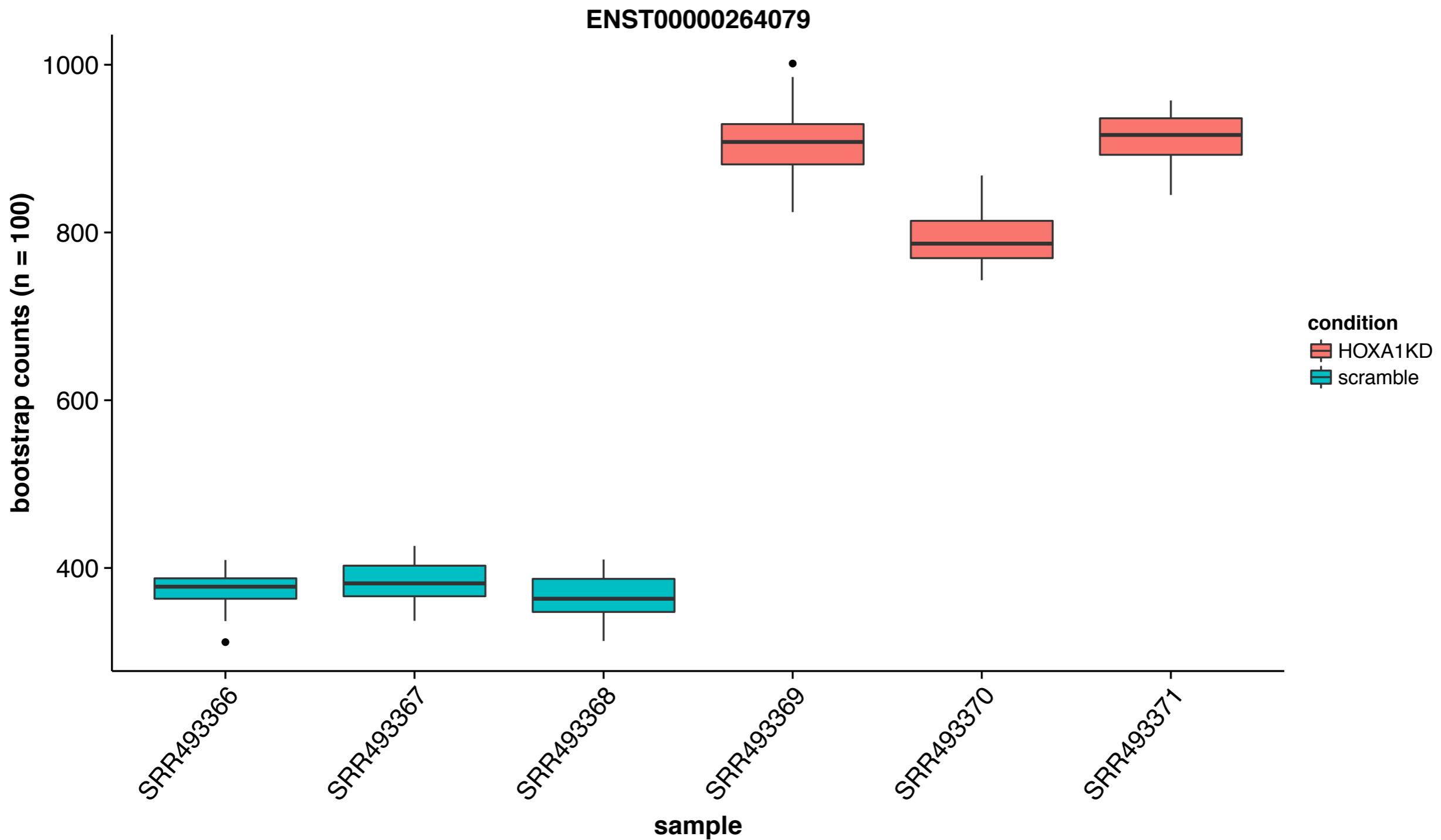
ENST00000265371
ENST00000374875
ENST00000374867
ENST00000413802
ENST00000395995

CLASP1

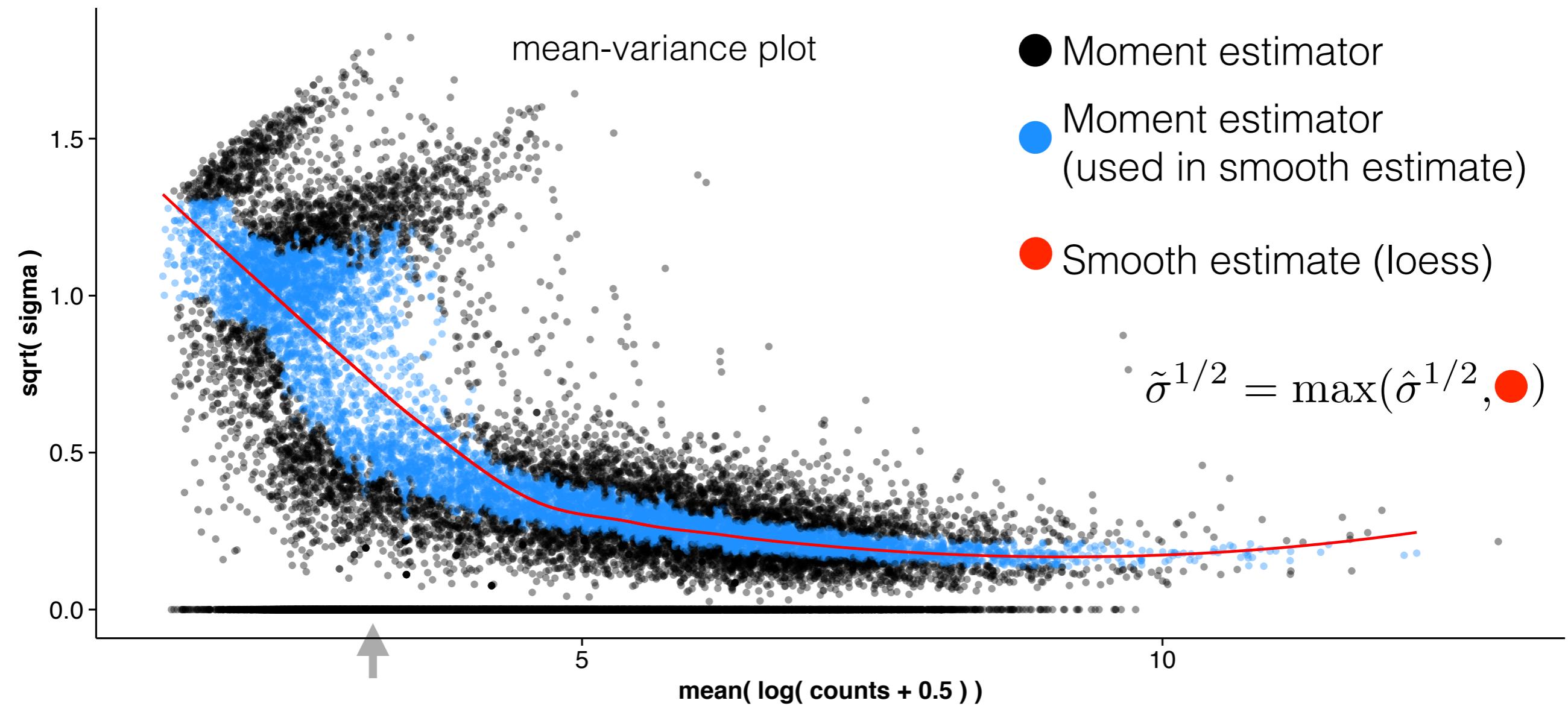




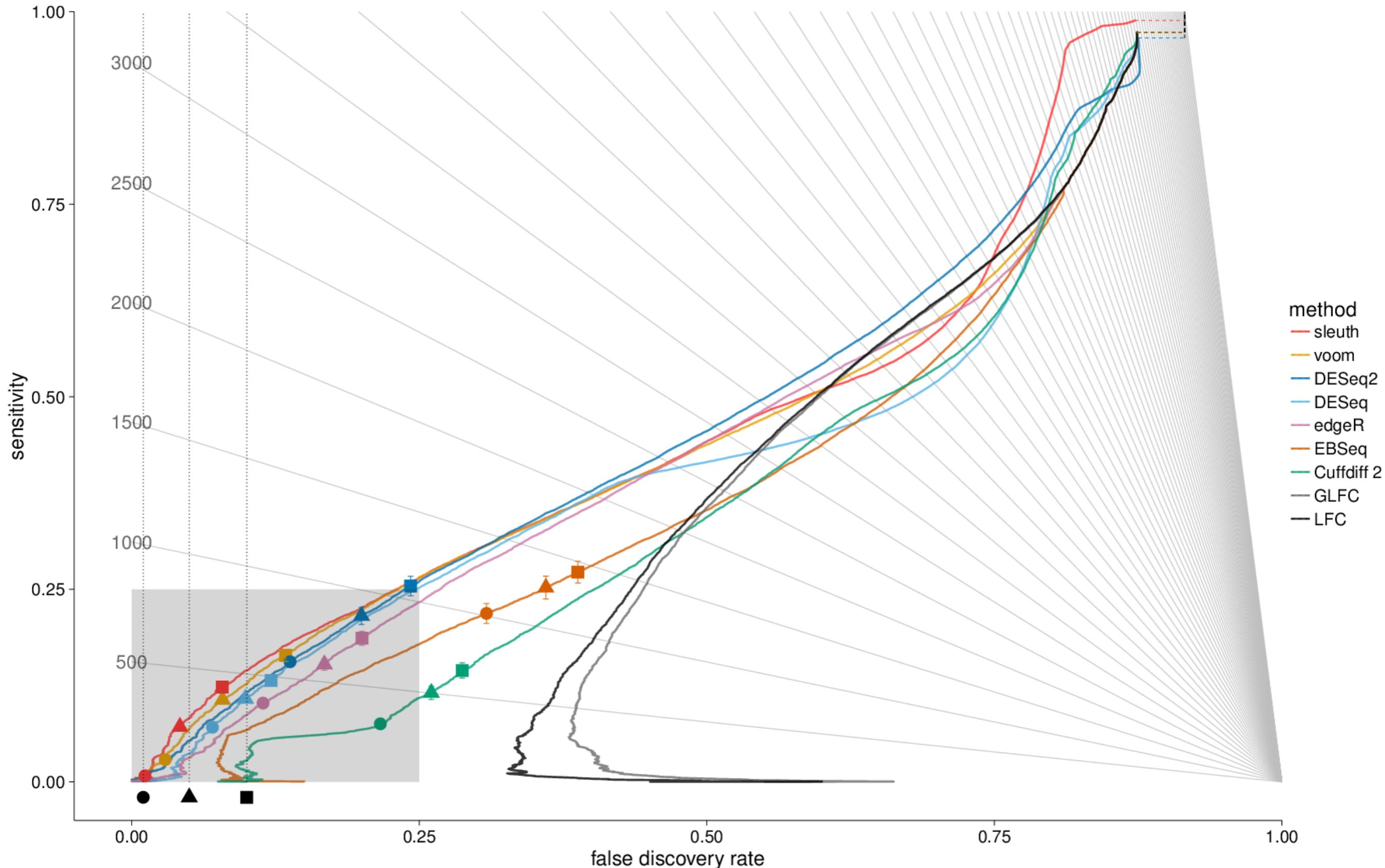
not significant by sleuth: FDR = 0.634
significant by DESeq2: FDR = 0.040

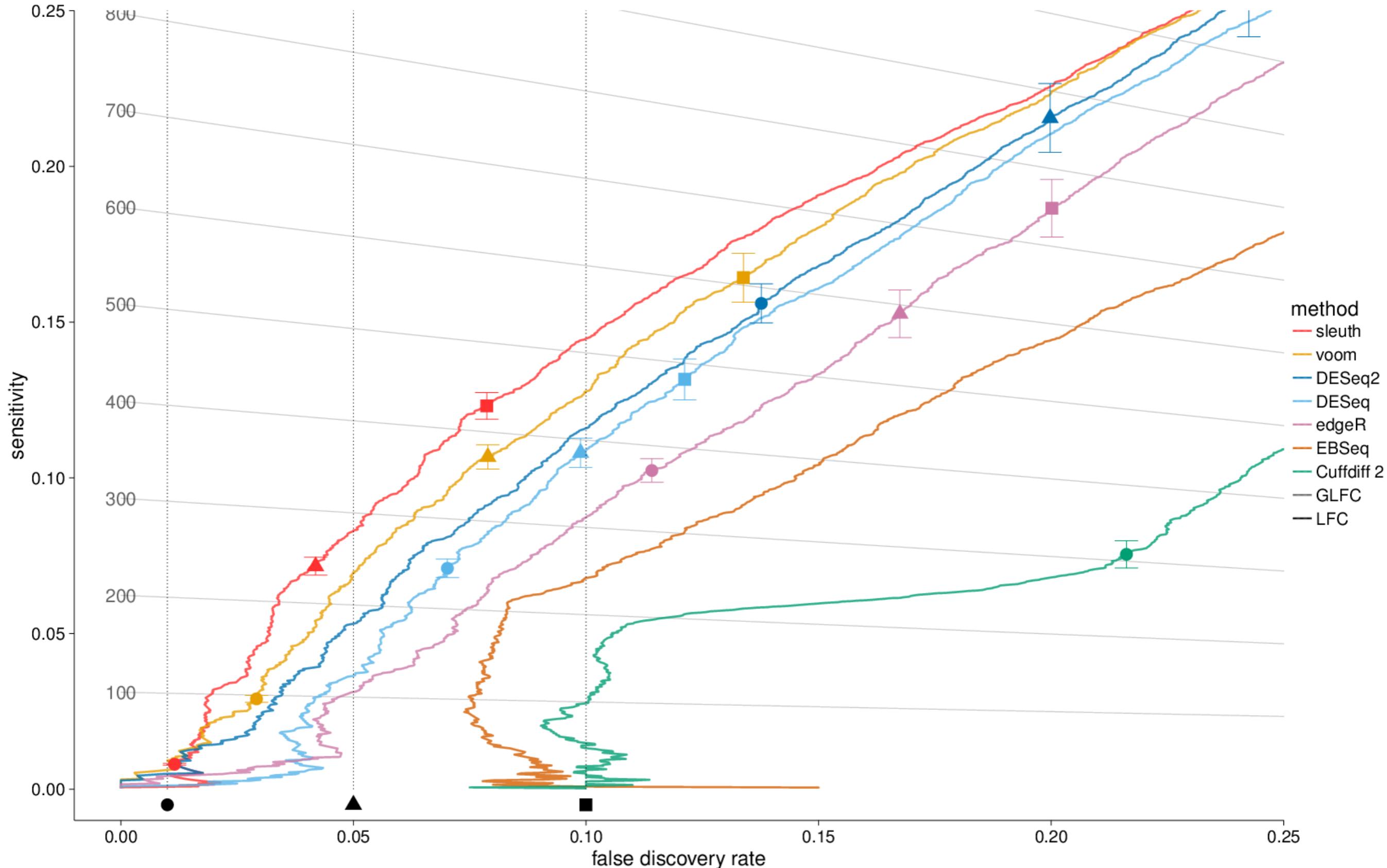


Biological variance estimation

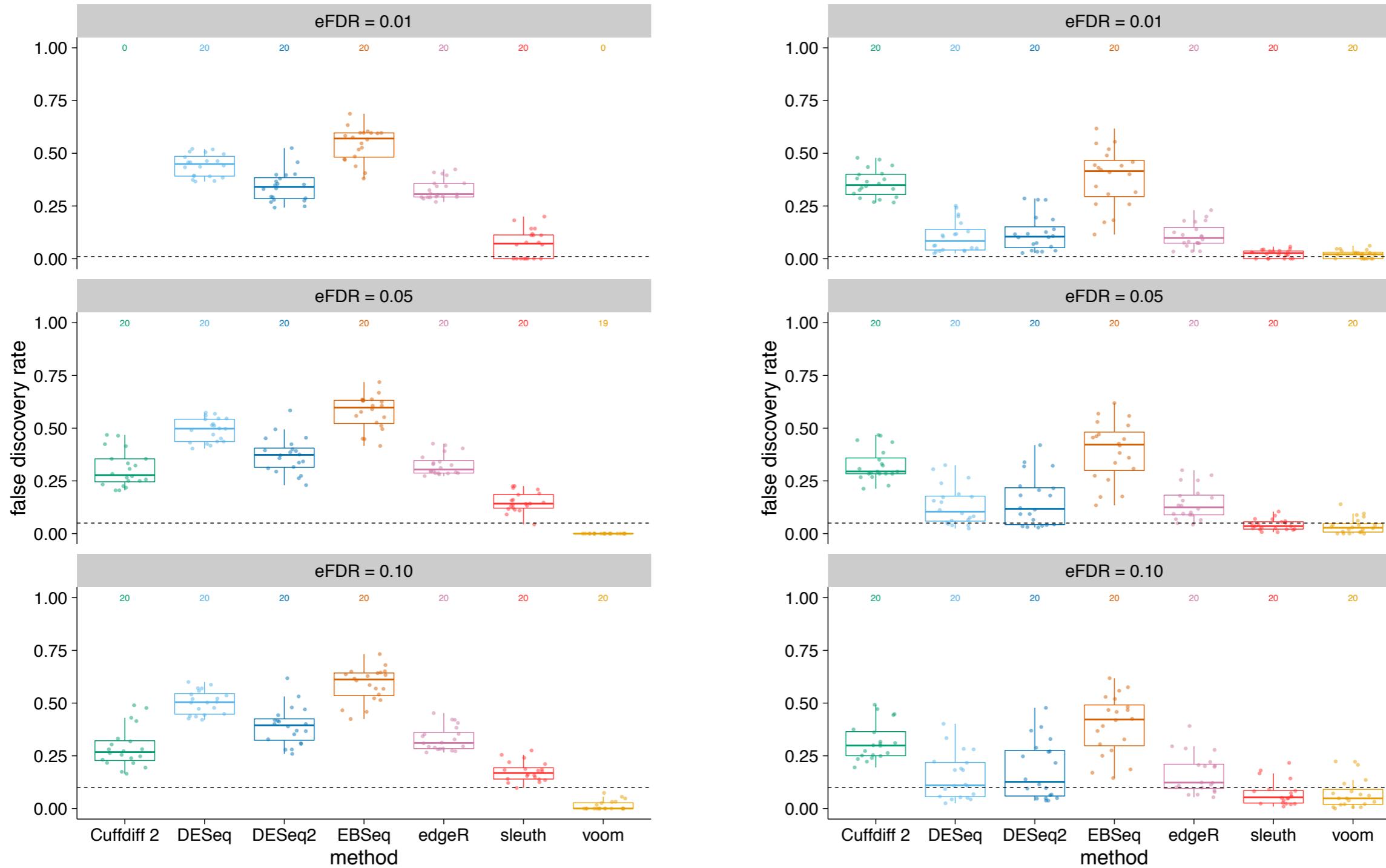


$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_i r_i^2 - \hat{\tau}^2$$





Negative control



Tissue specific expression

