Introduction to NGS data: experimental design

Matteo Fumagalli

Experimental design

You discovered a new species!

Experimental design

You have access to many samples!

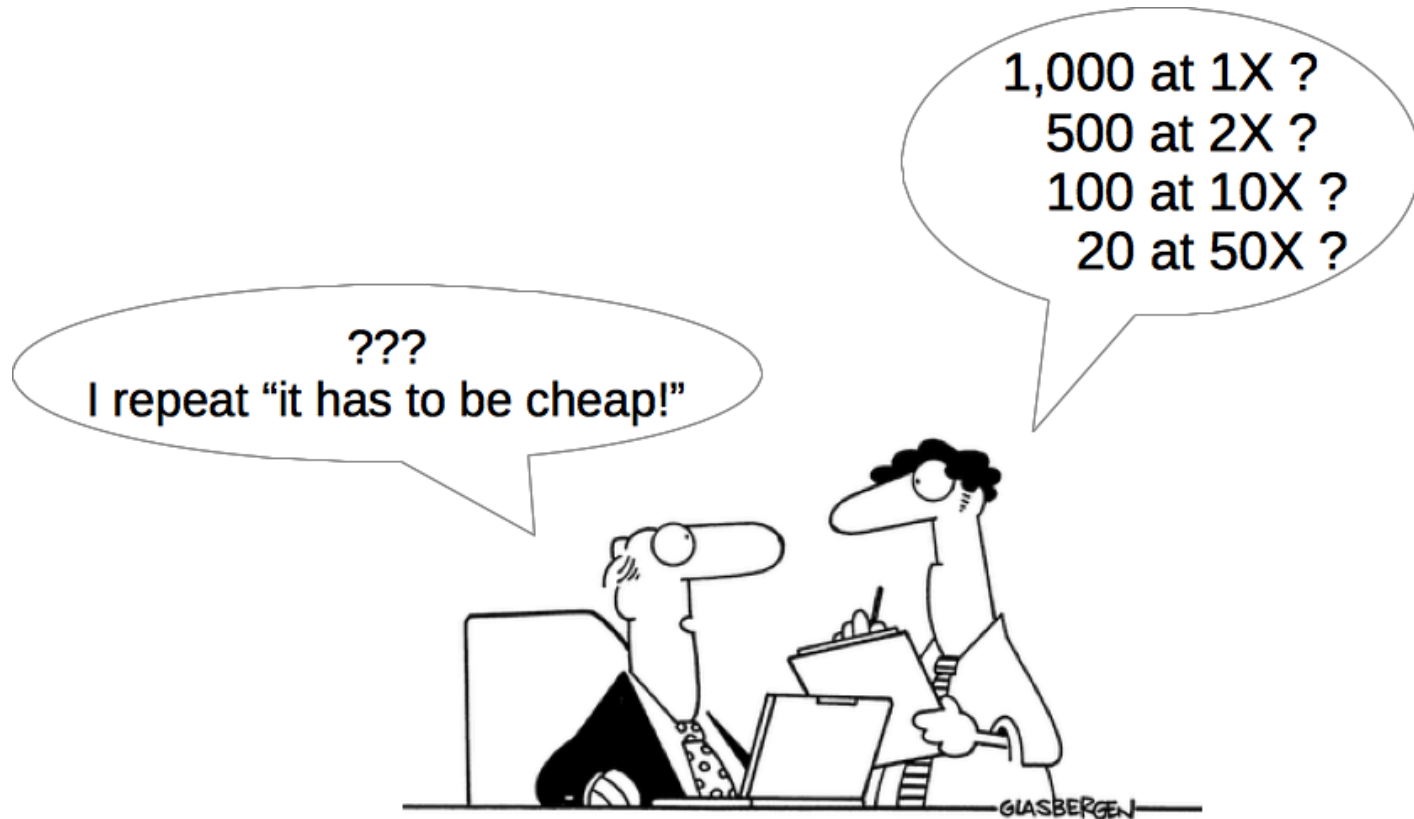# Experimental design

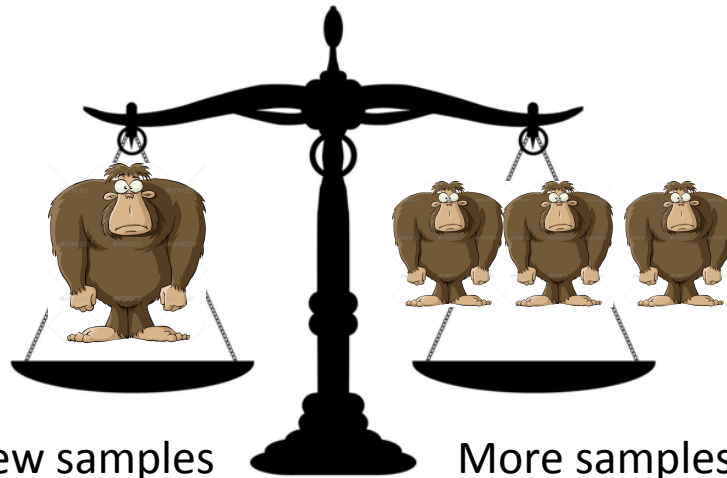# Experimental design

Experimental design

Experimental design

Experimental design

At a fixed budget…

Few samples
at higher quality

More samples
at lower quality

Simulations design

The sequencing strategy can easily be modeled in terms of the number of sequenced samples and the per-sample sequencing depth.

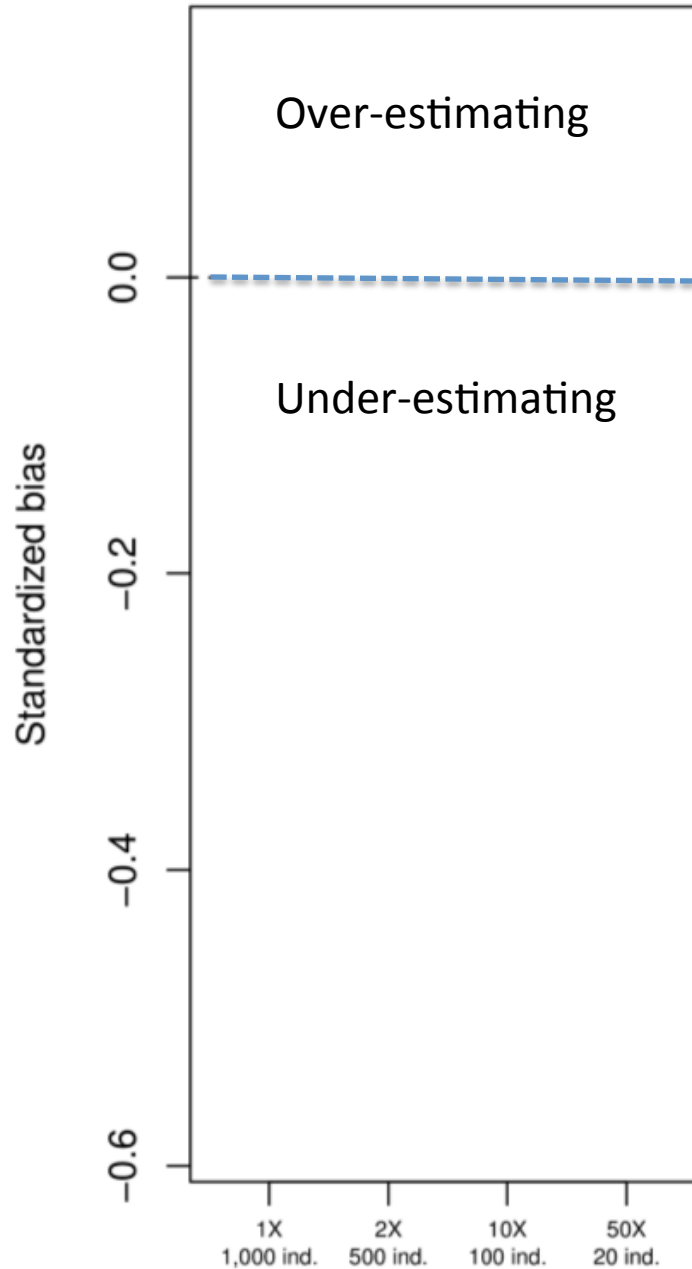| Sample size	Per-sample depth
1,000	1X
500	2X
100	10X
20	50X

total depth is
1,000X

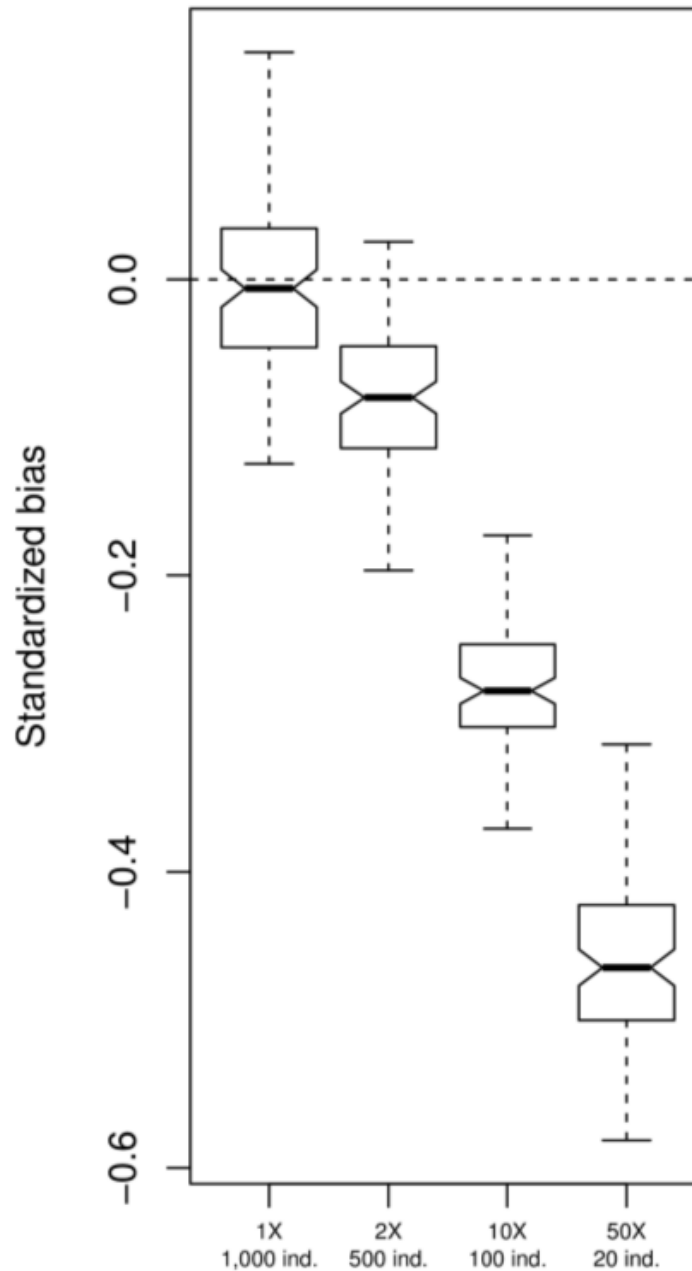# Number of segregating sites

$$Bias(S) = \frac{\hat{S} - S}{S}$$

Compared against the whole **population!**

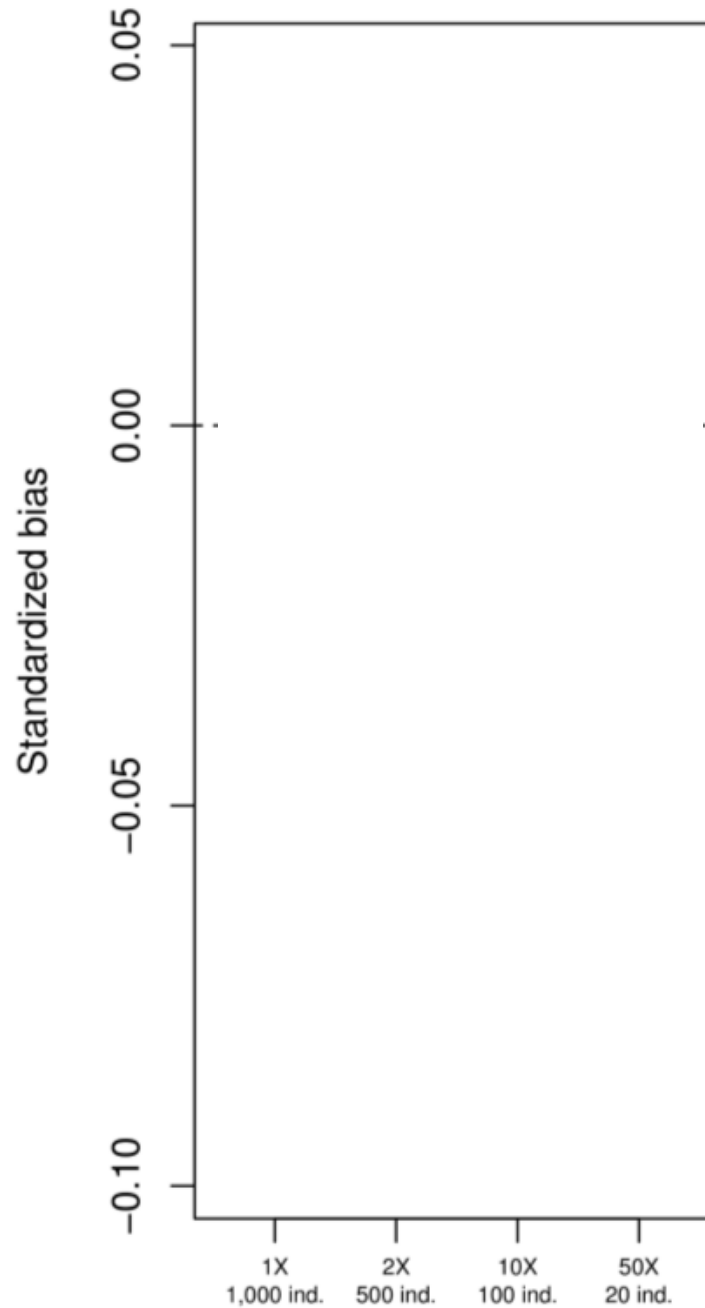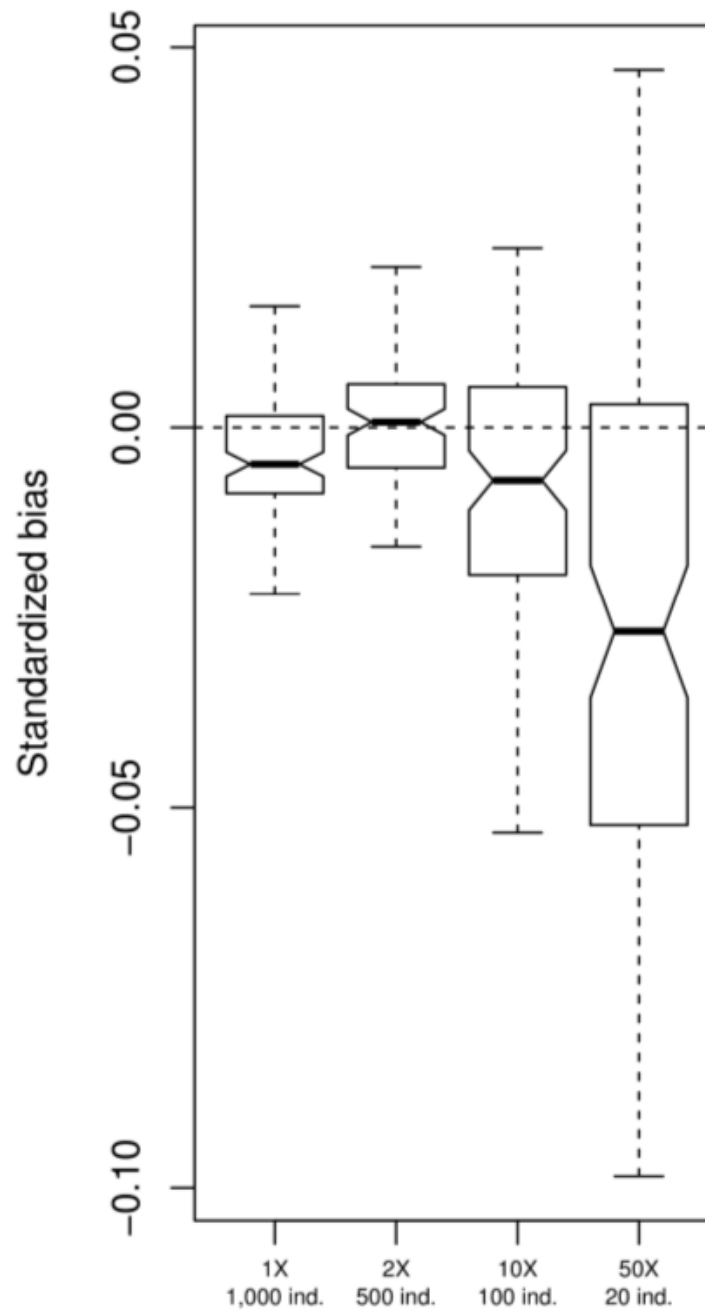# Number of segregating sites

$$Bias(S) = \frac{\hat{S} - S}{S}$$

Compared against the
whole **population!**

Standardized bias

0.0

−0.2

−0.4

−0.6

1X	2X	10X	50X
1,000 ind.	500 ind.	100 ind.	20 ind.

Expected heterozygosity

$$Bias(S) = \frac{\hat{S} - S}{S}$$

Compared against the
whole **population!**

Standardized bias

0.05

0.00

−0.05

−0.10

| 1X | 2X | 10X | 50X |
| 1,000 ind. | 500 ind. | 100 ind. | 20 ind. |

Expected heterozygosity

$$Bias(S) = \frac{\hat{S} - S}{S}$$

Compared against the whole **population!**

Standardized bias

1X	2X	10X	50X
1,000 ind.	500 ind.	100 ind.	20 ind.

SNP calling

False Positive

Rate

0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00

| 1X | 2X | 10X | 50X |
| 1,000 ind. | 500 ind. | 100 ind. | 20 ind. |

False Negative

Rate

0.5
0.4
0.3
0.2

| 1X | 2X | 10X | 50X |
| 1,000 ind. | 500 ind. | 100 ind. | 20 ind. |

False Positive   **SNP calling**   **False Negative**

Rate

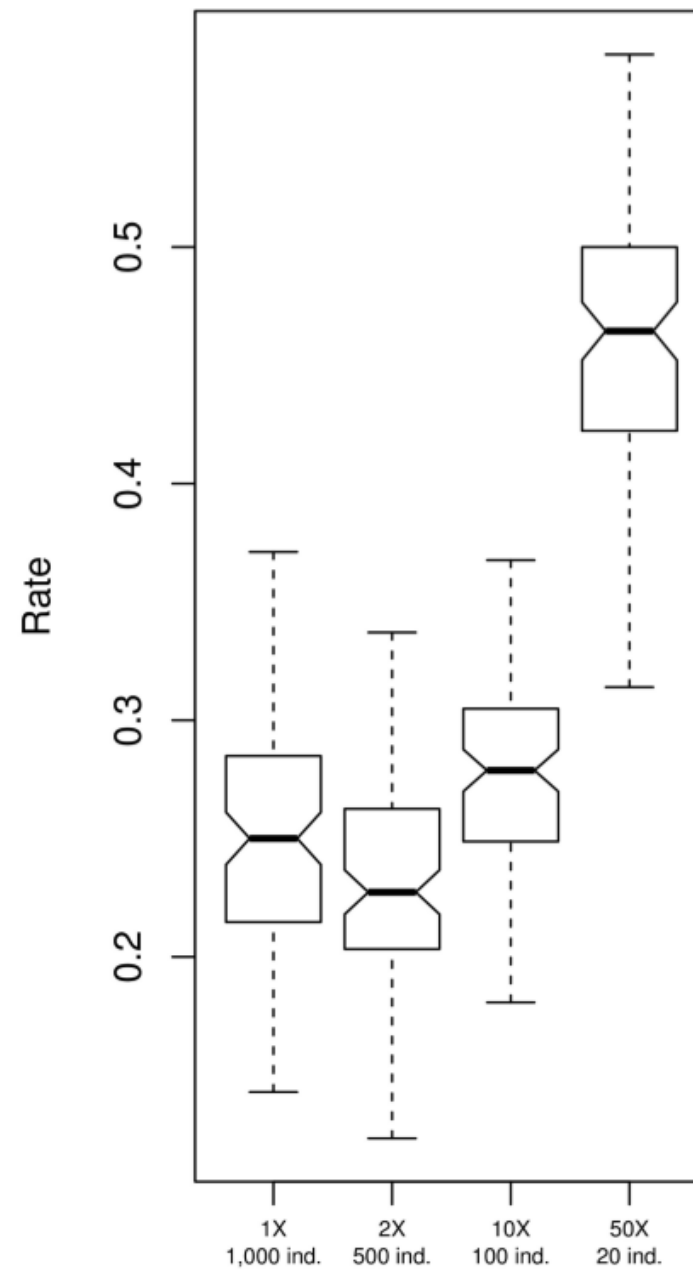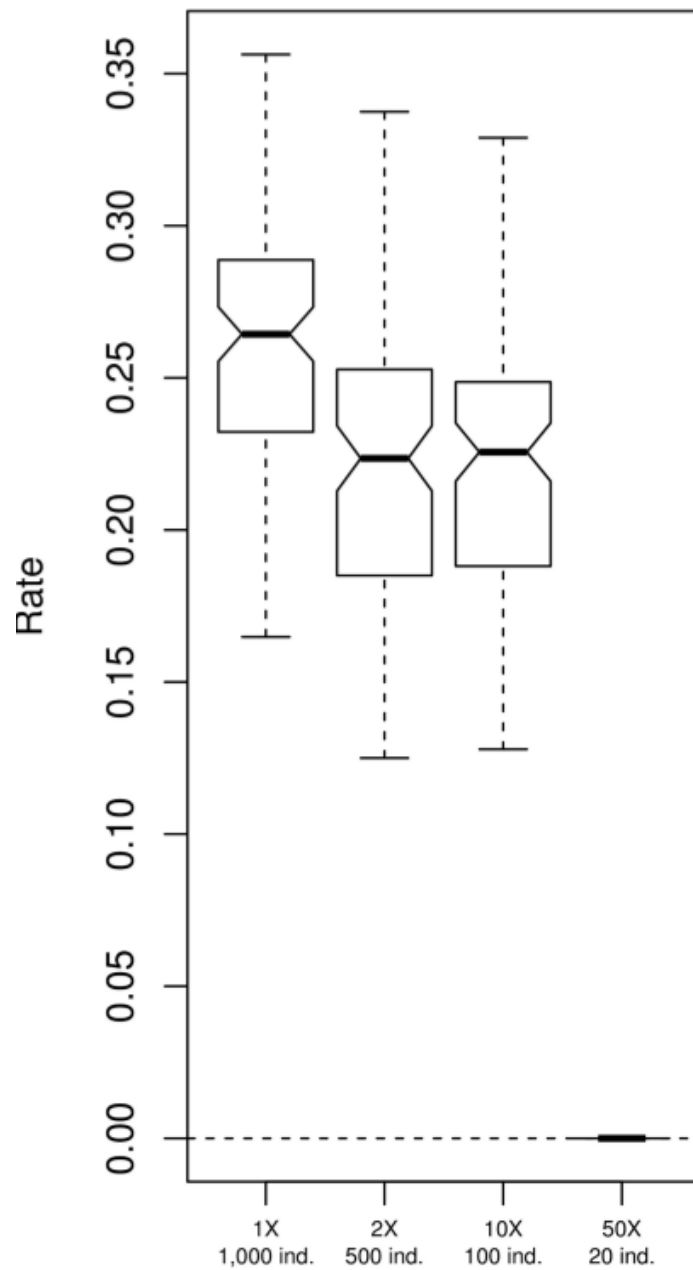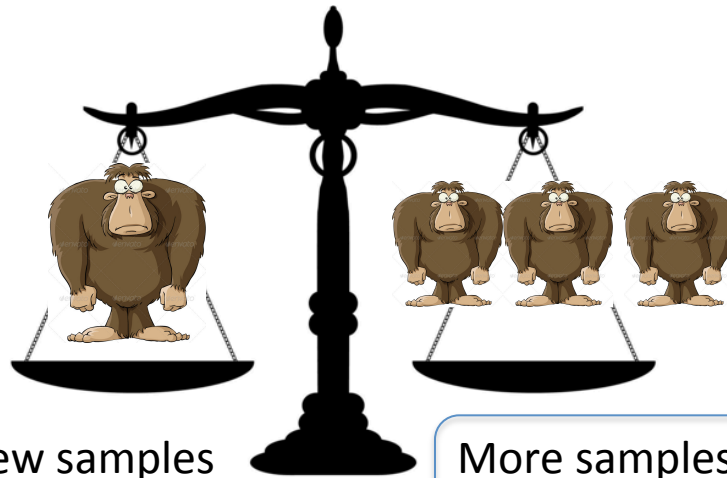| 1X | 2X | 10X | 50X |
| 1,000 ind. | 500 ind. | 100 ind. | 20 ind. |
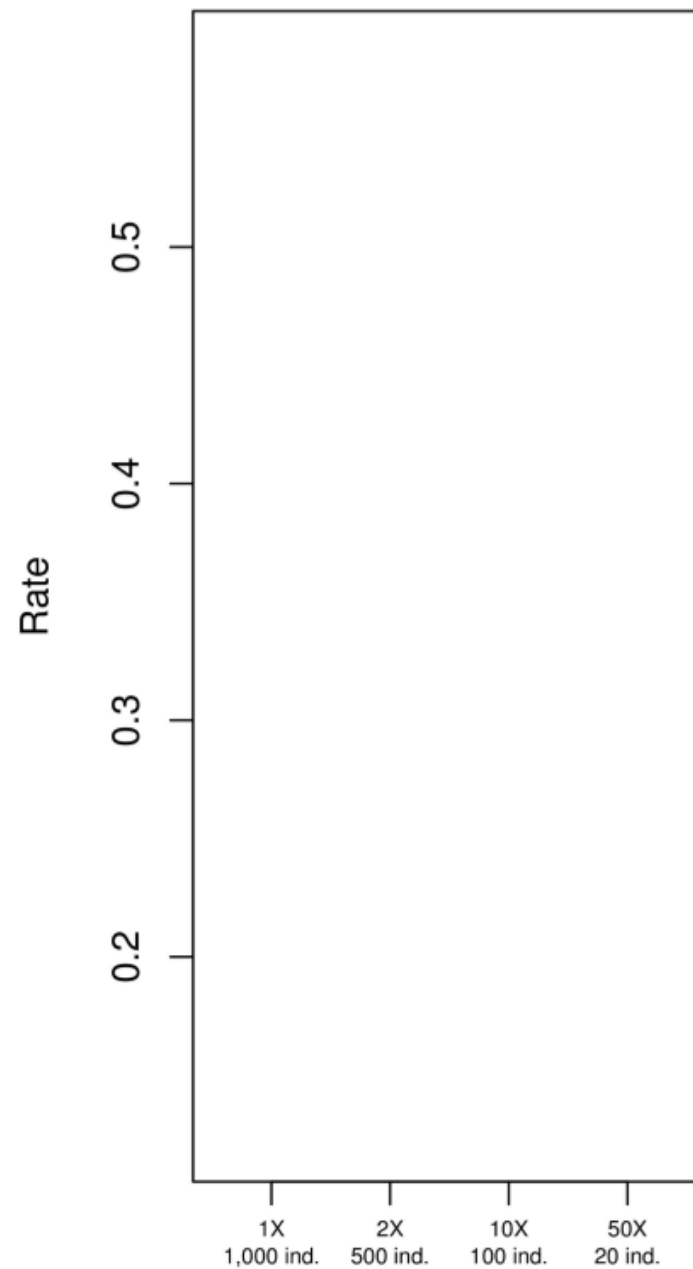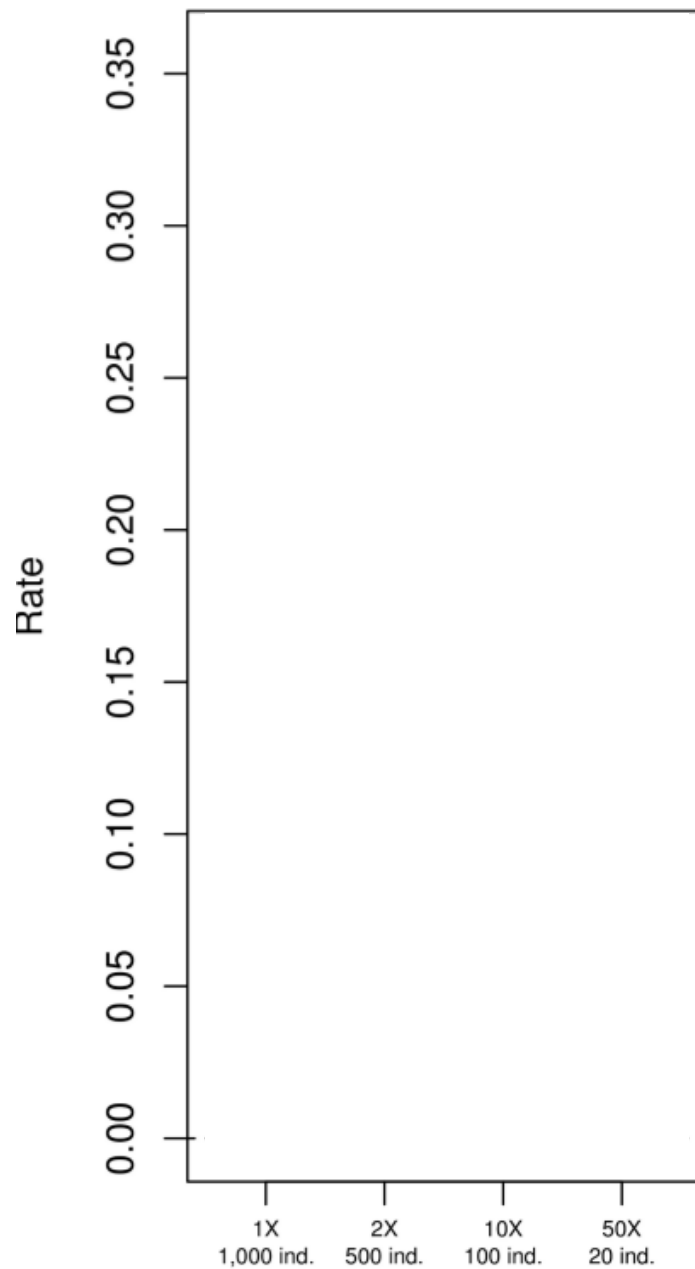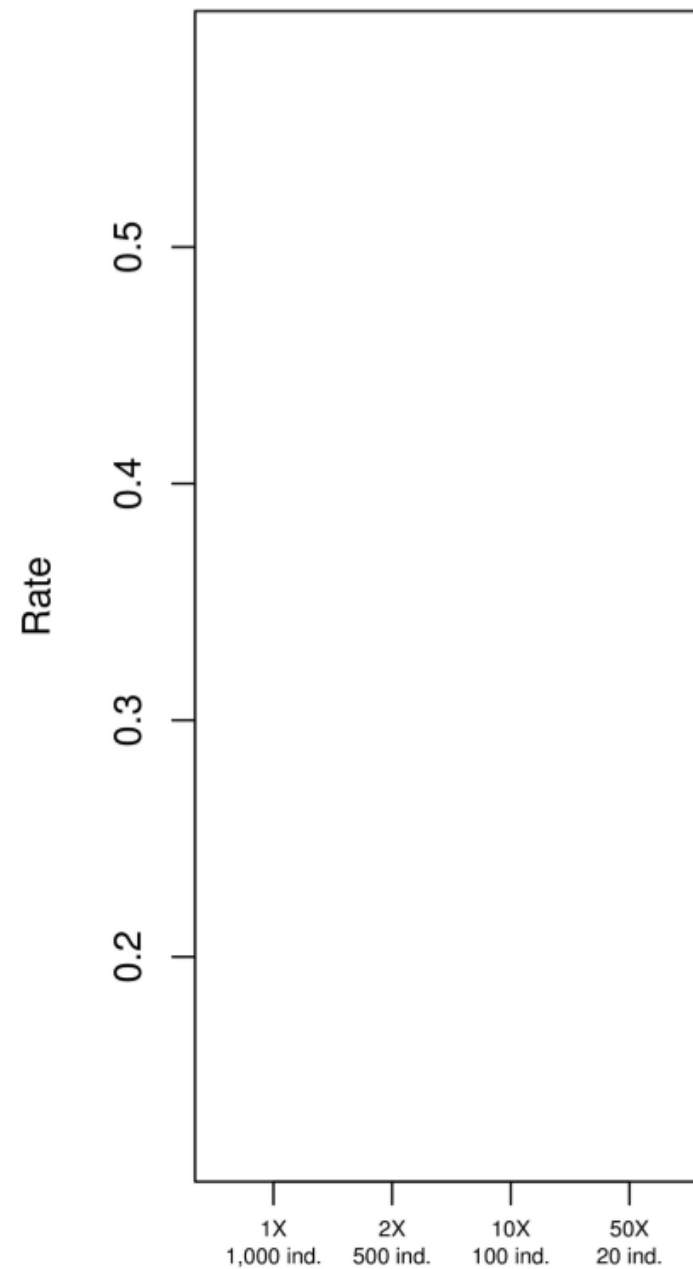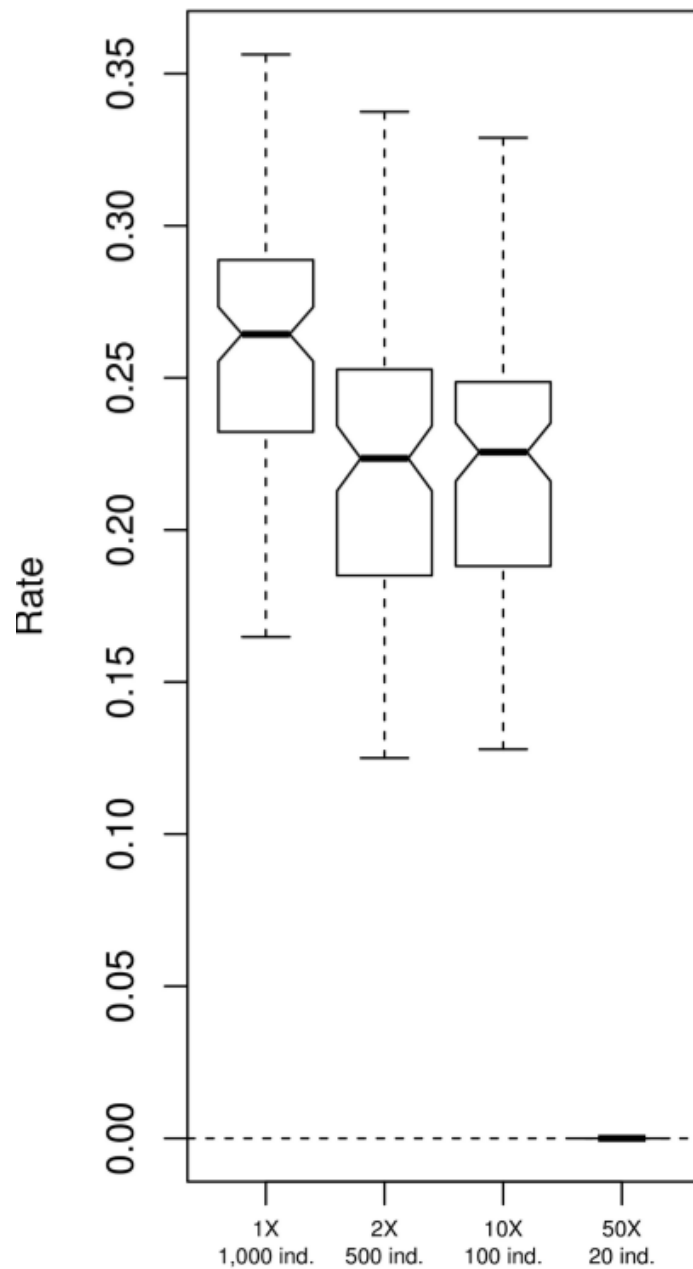
False Positive   **SNP calling**   **False Negative**

Experimental design

At a fixed budget…

Few samples
at higher quality

More samples
at lower quality

For detecting selection:
Sample allele frequencies will be a better representation of population variation