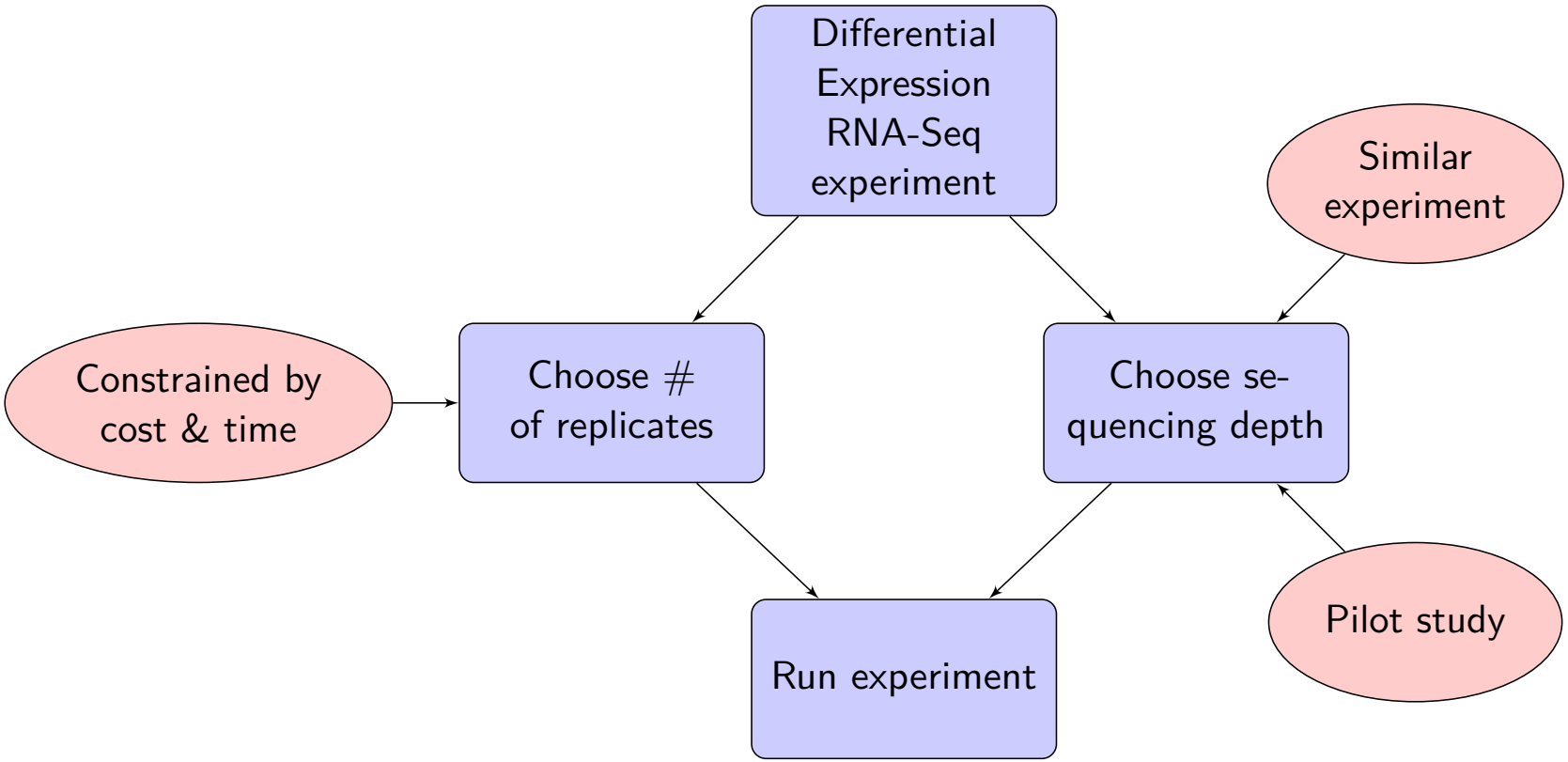
Predicting read depth to maximize statistical power in RNA-Seq experiments

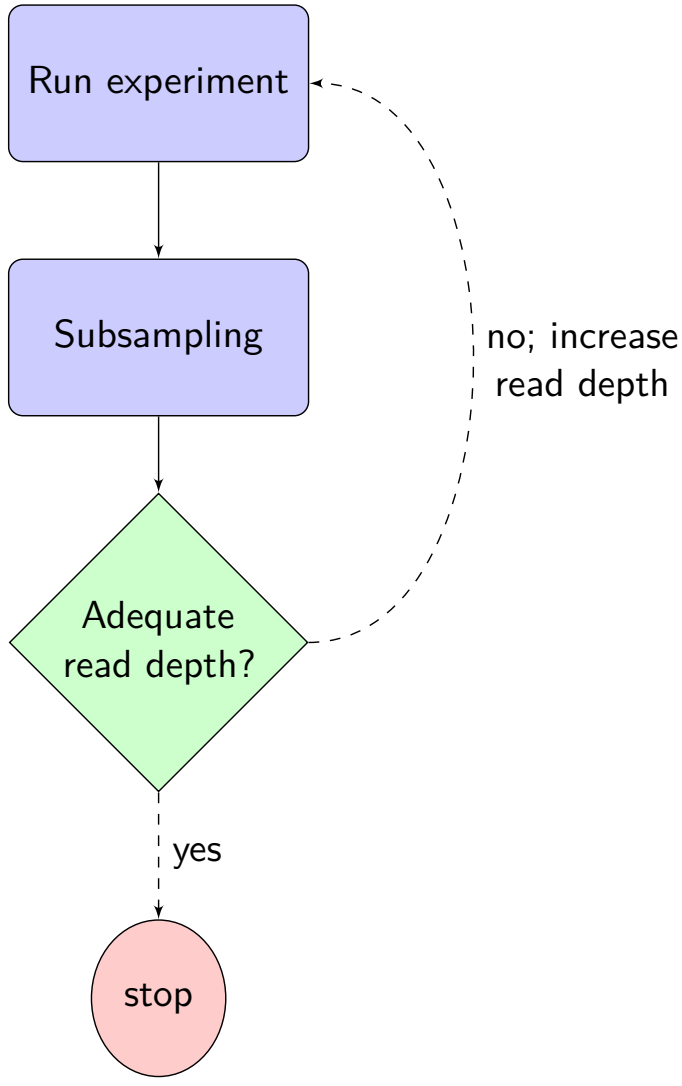
Andrew Bass
Princeton University

November 18, 2015

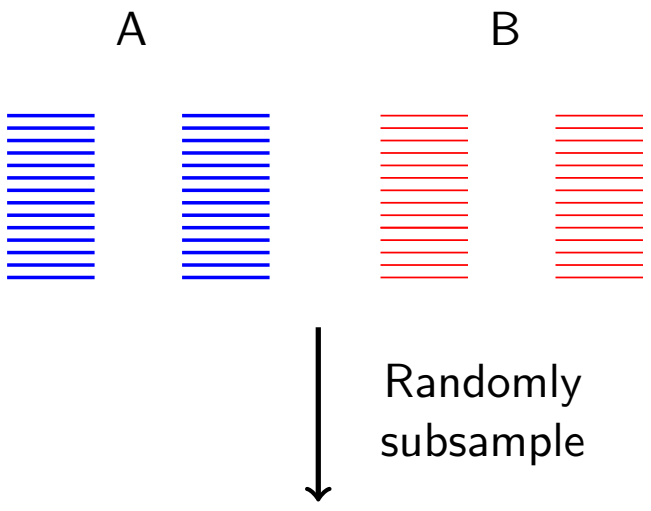
Designing an experiment



Determining the appropriate read depth



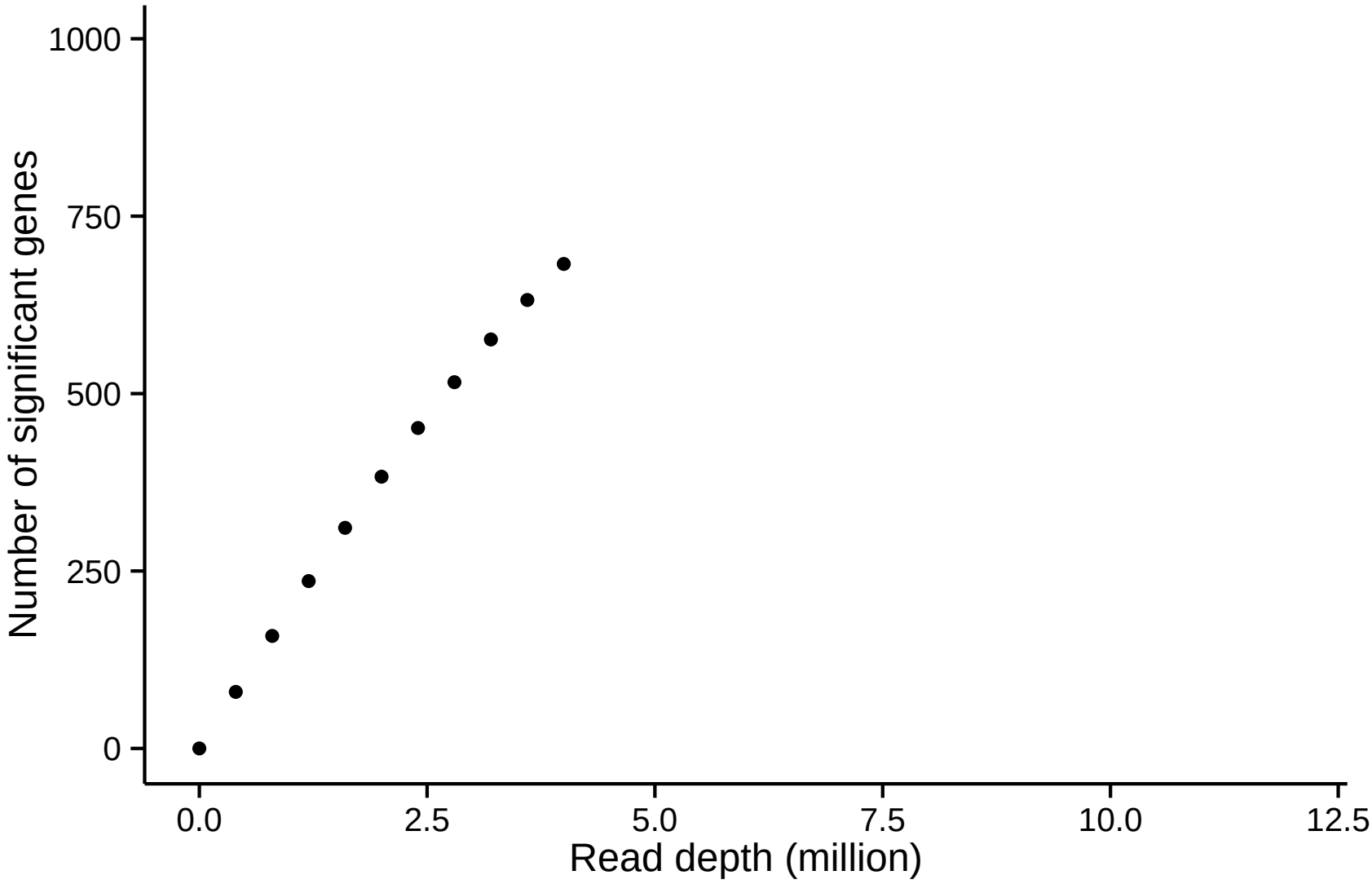
Subsampling



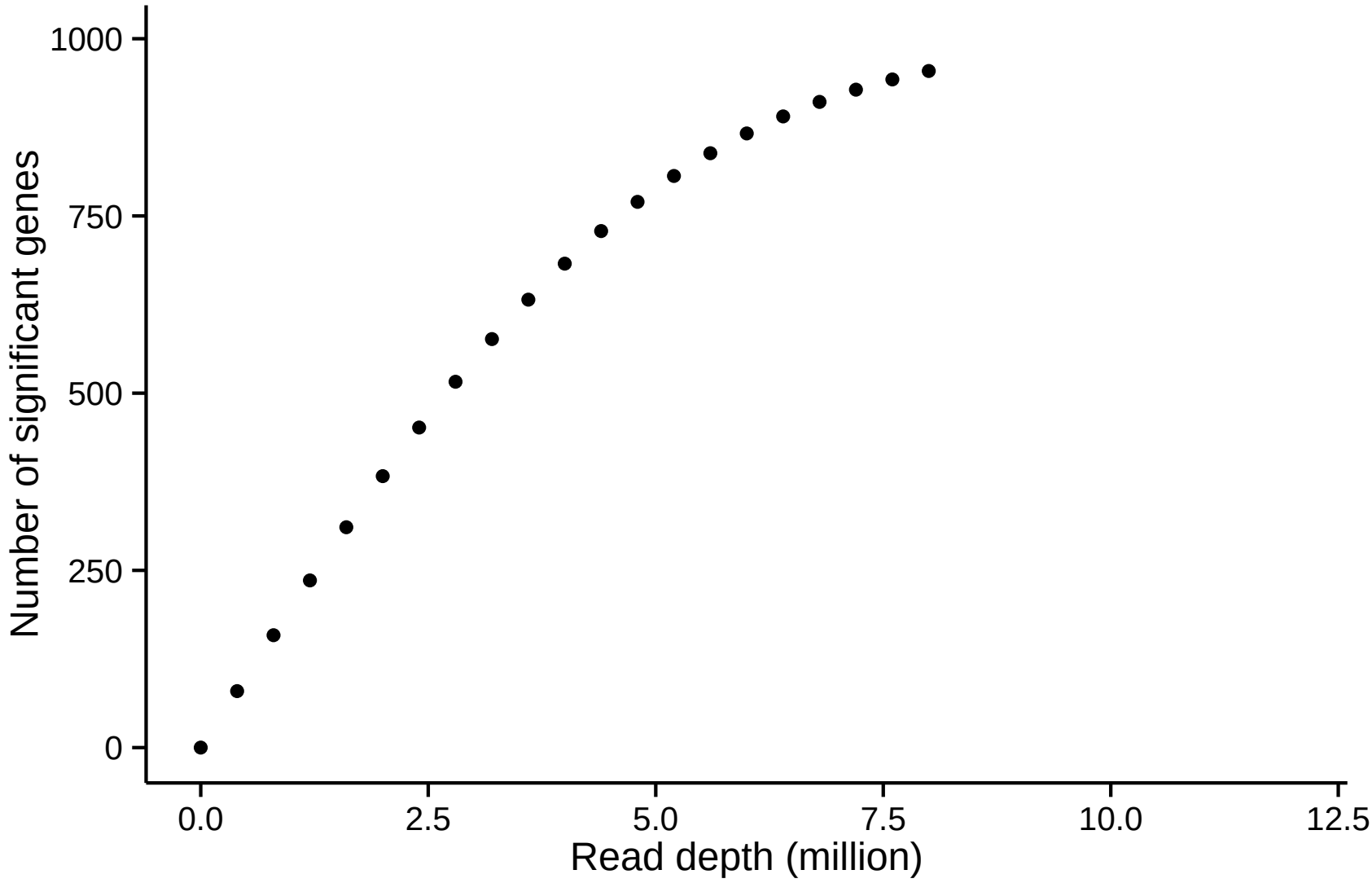
How many differentially expressed genes?

How many differentially expressed genes?

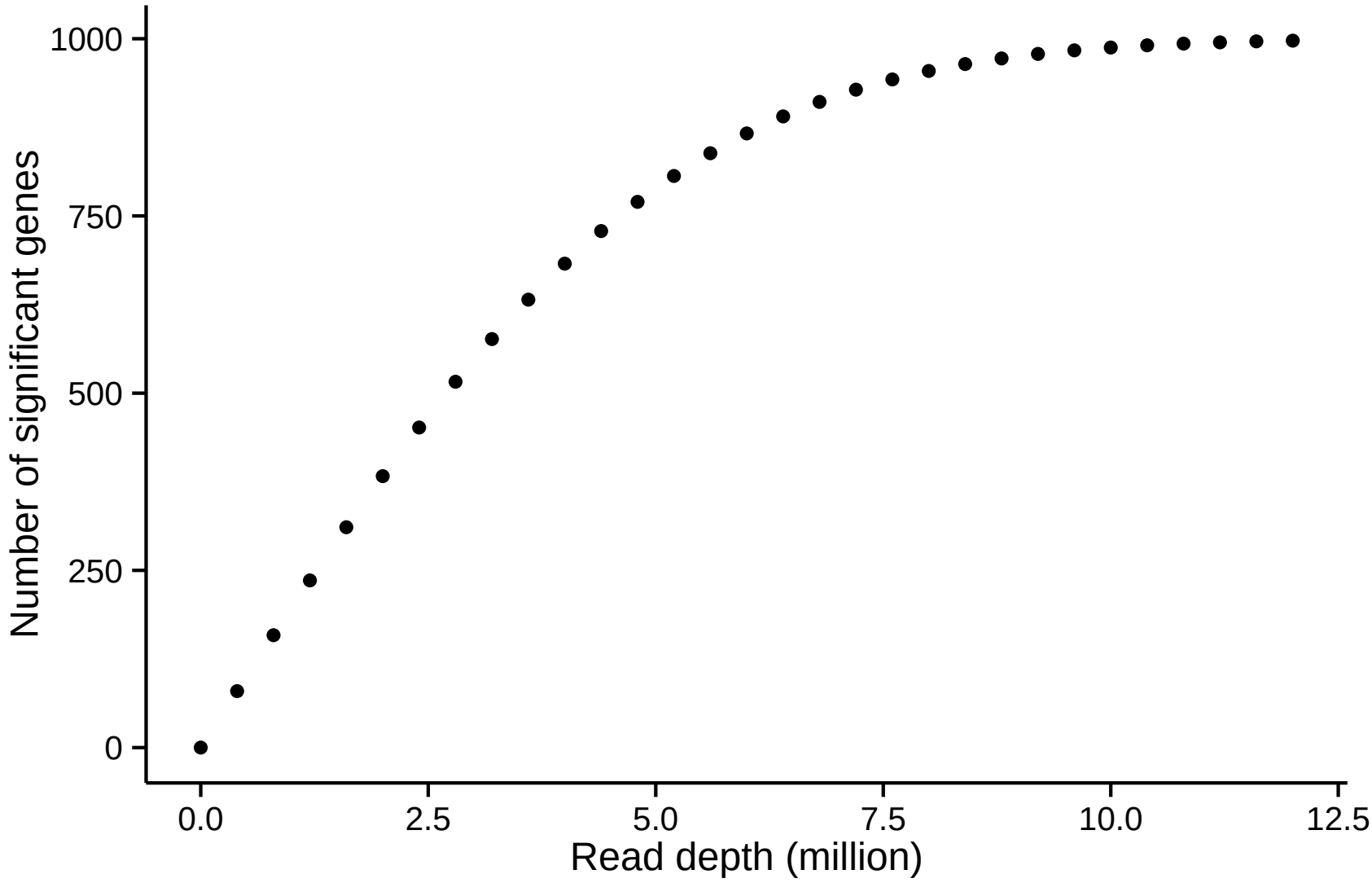
Is the power saturated at current read depth?



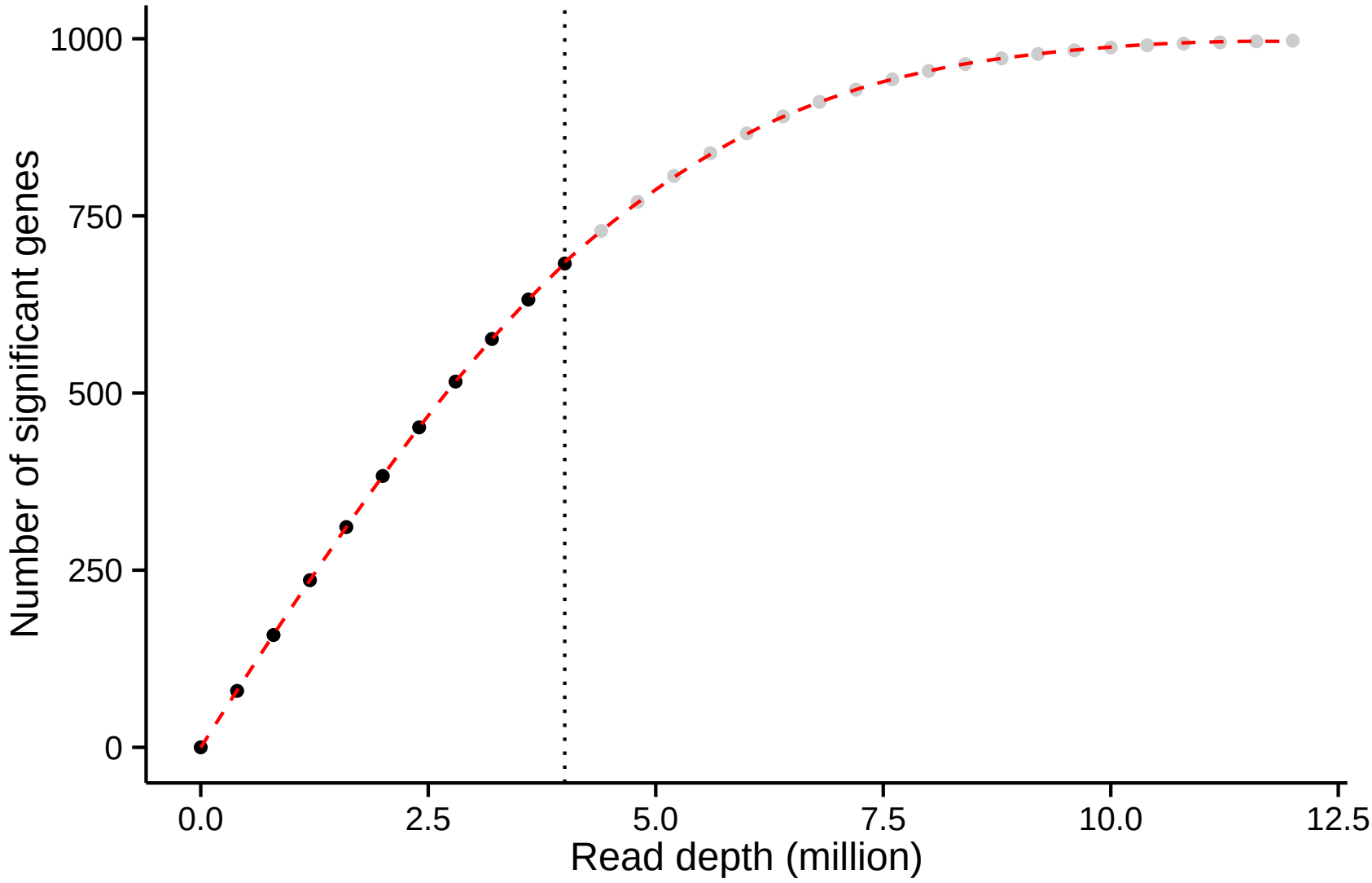
Is the power saturated at current read depth?



Is the power saturated at current read depth?



Can we model saturation curves?



Per-gene power

$$\alpha_i(p) = \begin{cases} 1, & i \in \Omega \text{ and } pD_i > \lambda \\ 0, & \text{otherwise.} \end{cases}$$

- $\alpha_i(p)$: indicator function that gene i is significant at p
- p: subsampling proportion
- Ω: set of significant genes
- D_i: read depth of gene i
- λ: minimum read depth in Ω

General model to utilize per-gene power

$$\mathbb{E}[G(p)] = \sum_{i=1}^{m} \mathbb{E}[\alpha_i(p)] \tag{1}$$

$$\mathbb{E}[G(p)] = m(1-\pi_0)\,\mathrm{Pr}\left(D > \frac{\lambda}{p} \,\middle|\, \Omega\right) \tag{2}$$

$$\mathbb{E}[G(p)] = m(1-\pi_0)\left[1 - F\left(\frac{\lambda}{p} \,\middle|\, \Omega\right)\right] \tag{3}$$

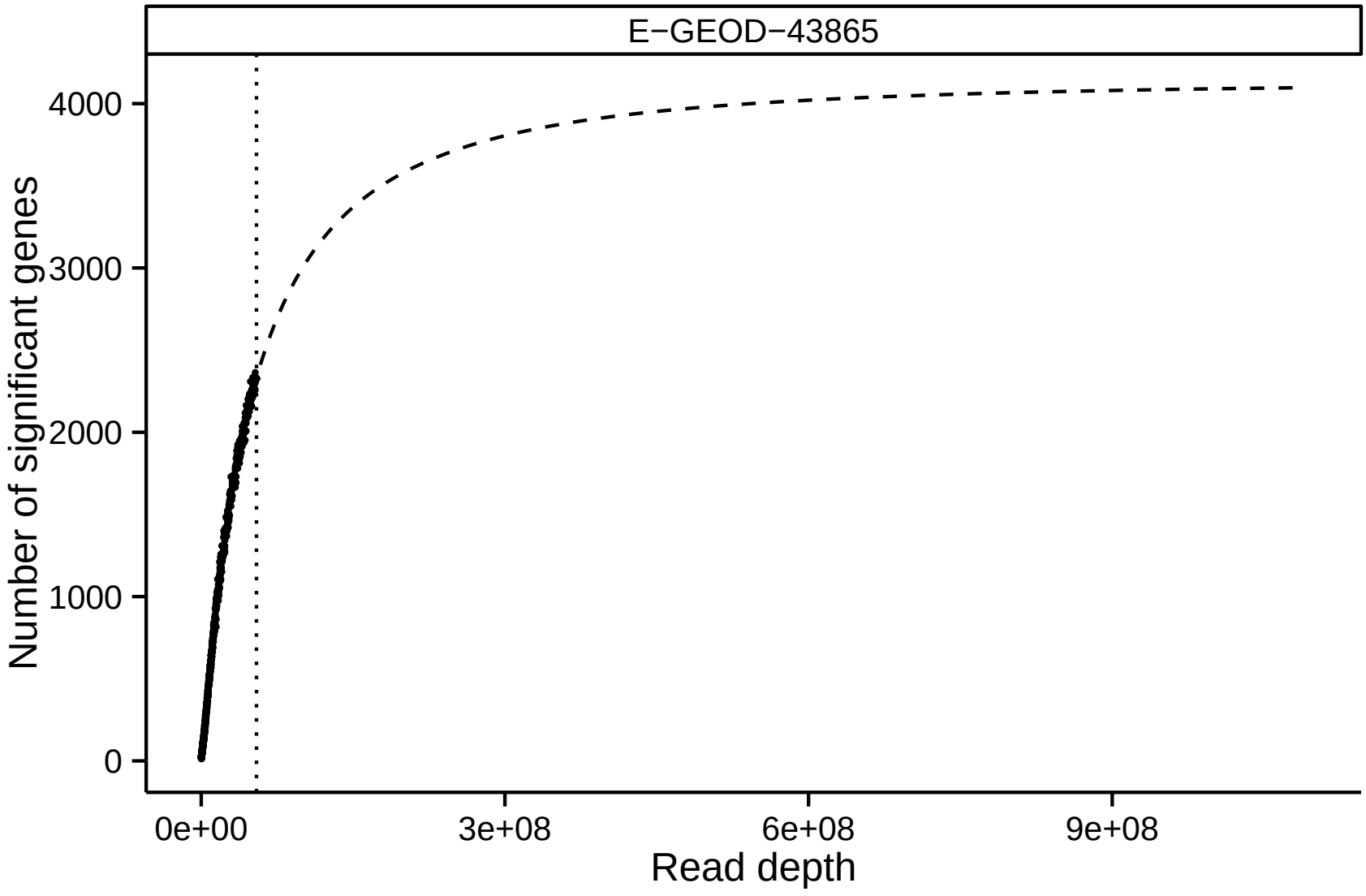
- $G(p)$: number of significant genes at p
- m: number of genes
- π_0: proportion of non-significant genes
- F: cumulative distribution function of per-gene read depth

Think log-normal!

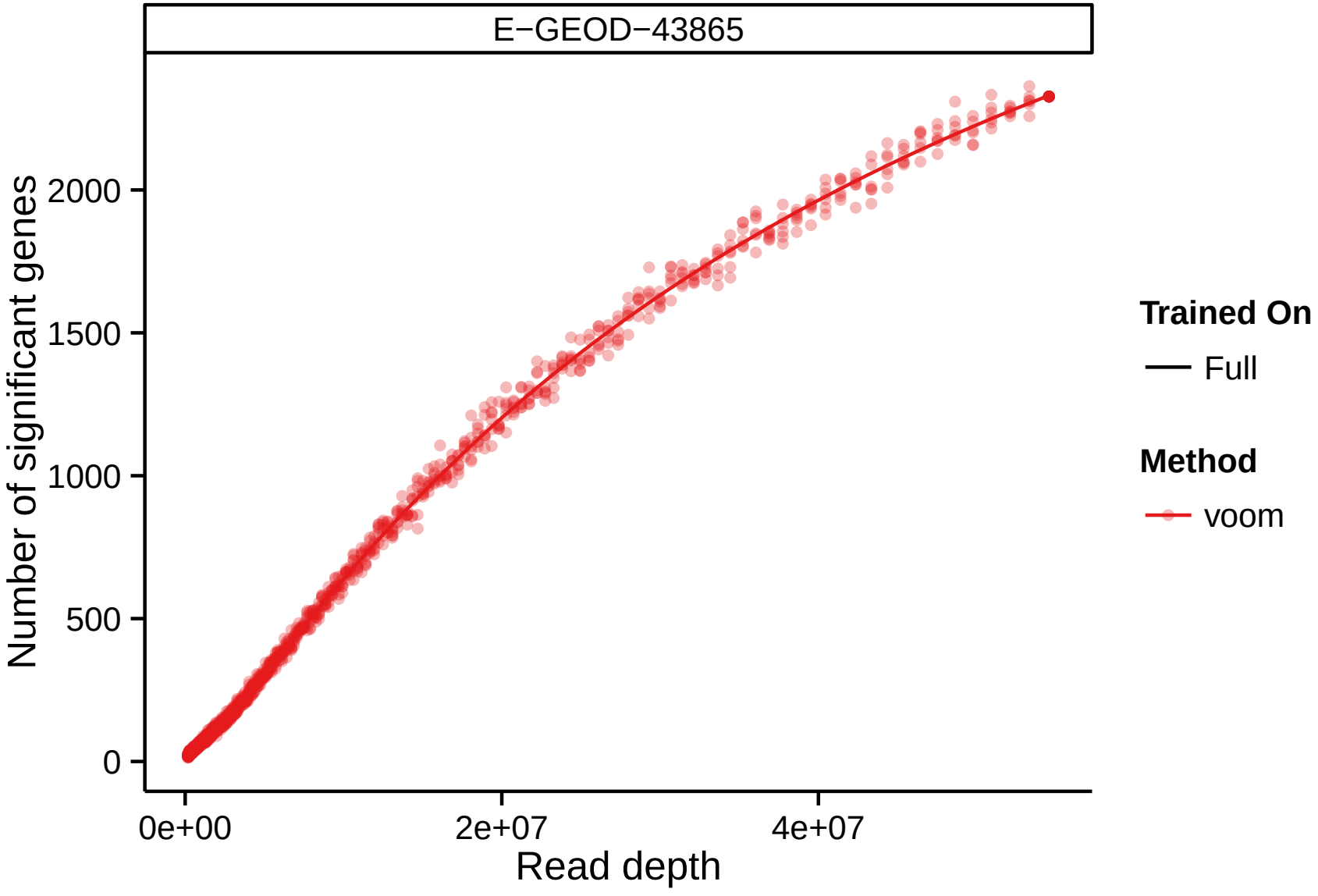
$$\mathbb{E}[G(p)] = m(1-\pi_0)\Phi\left(\frac{log(p+b)-\mu^*}{\sigma}\right)$$

- $m(1-\pi_0)$: number of significant genes in the experiment
- Φ: Normal CDF
- μ^*, σ: adjusted mean and standard deviation from the distribution of counts
- p: subsampling proportion
- b: offset

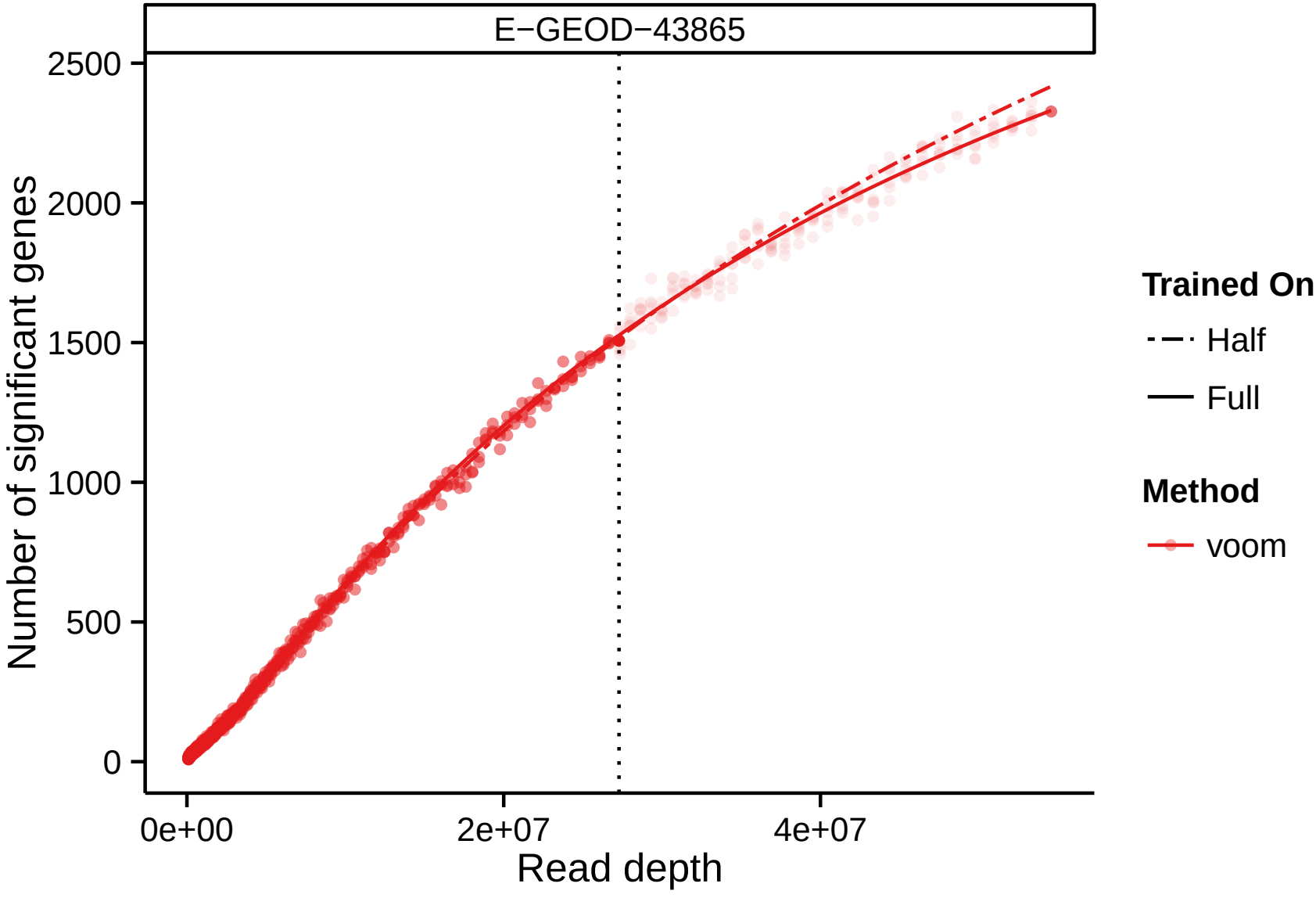
How do you use the methodology?



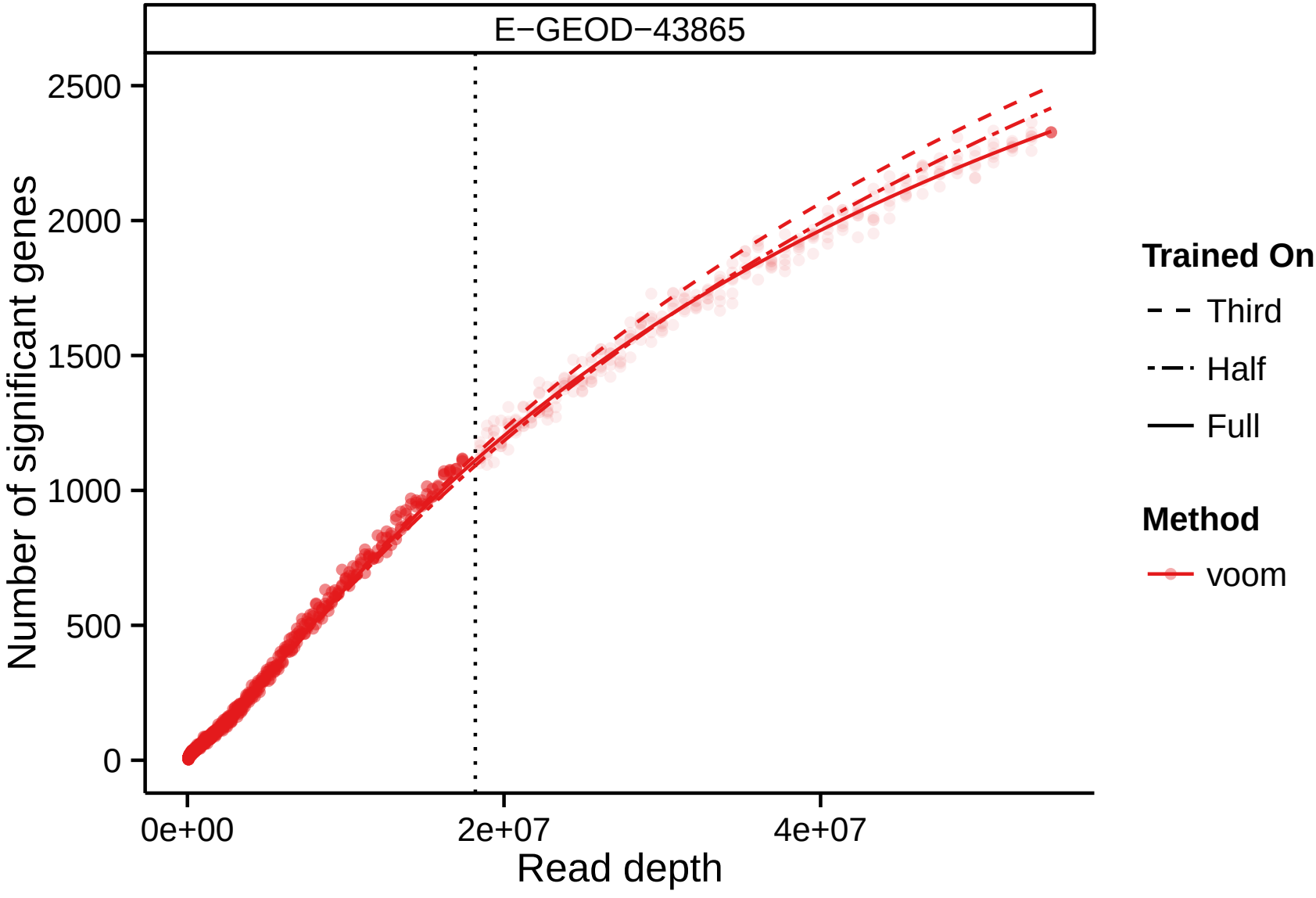
Assessing the model fits



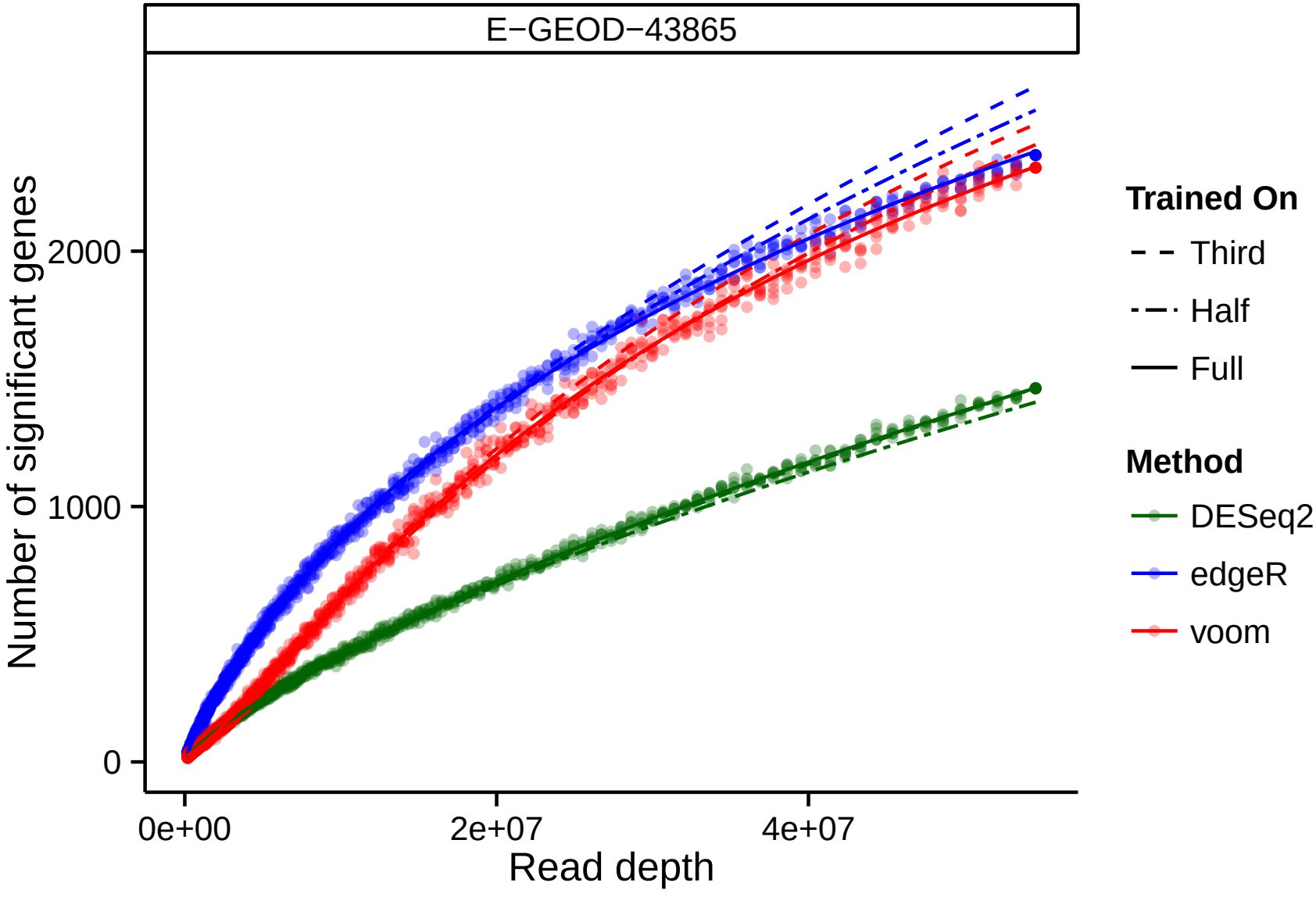
Assessing the model fits



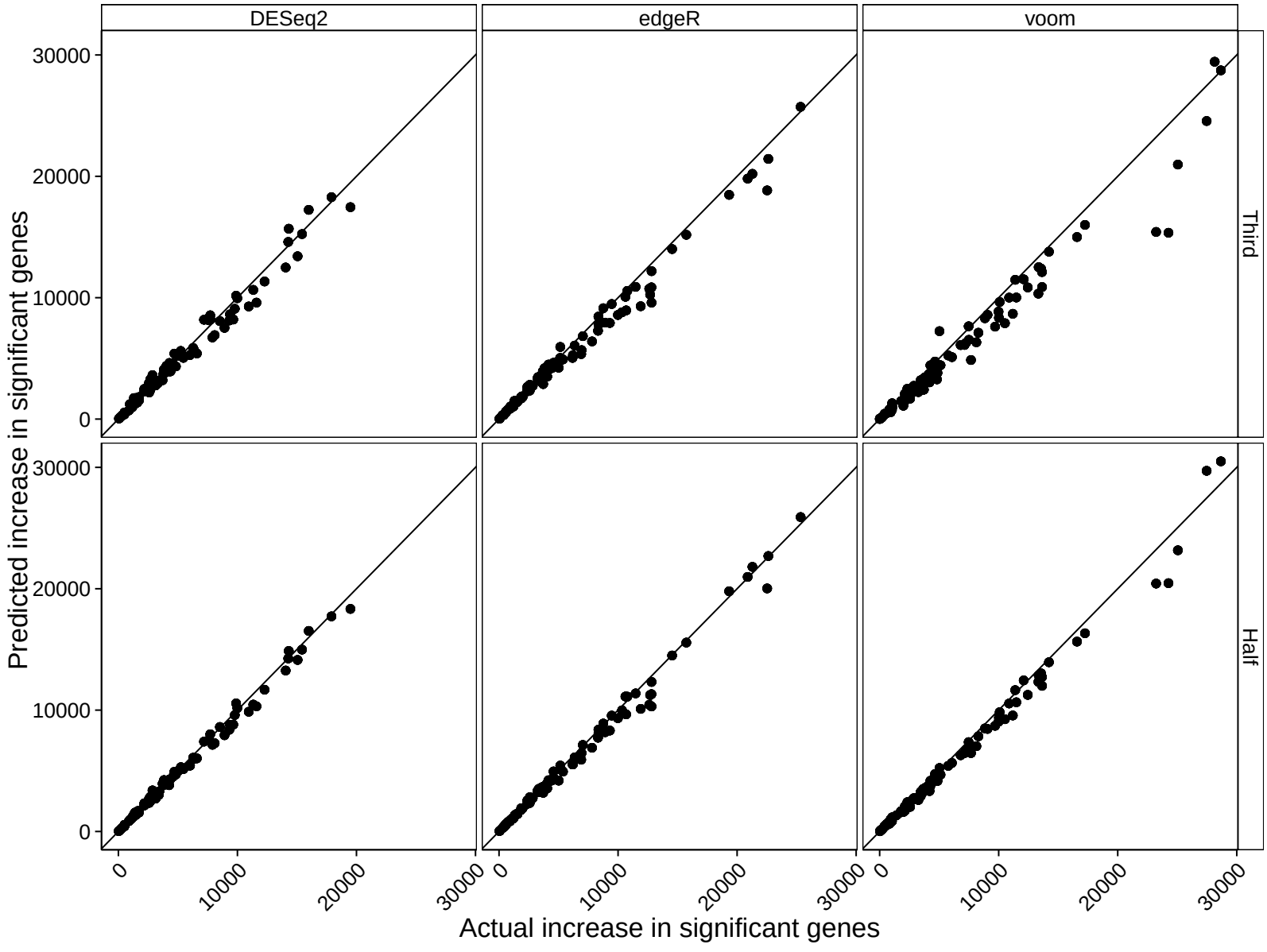
Assessing the model fits



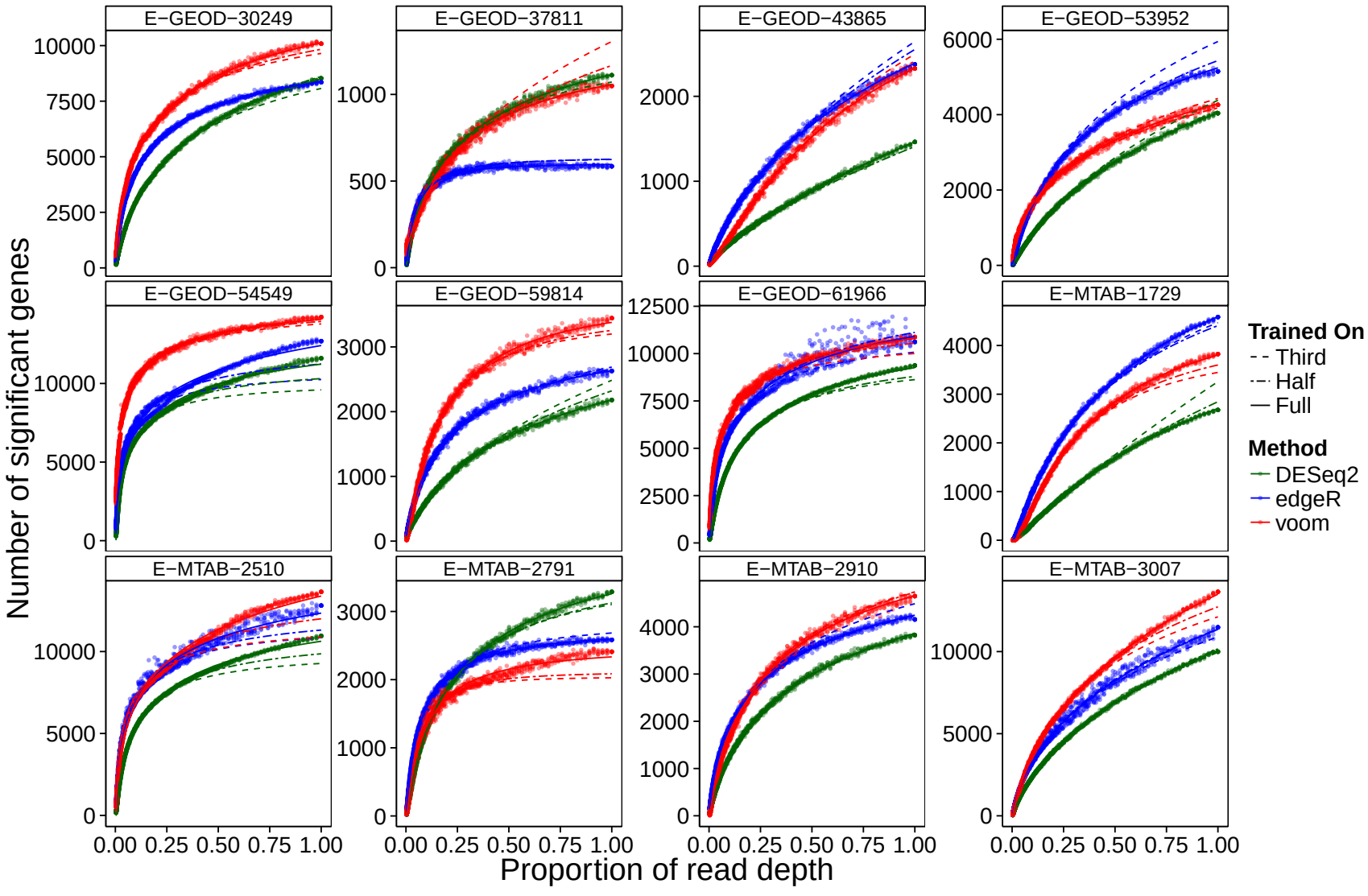
Assessing the model fits



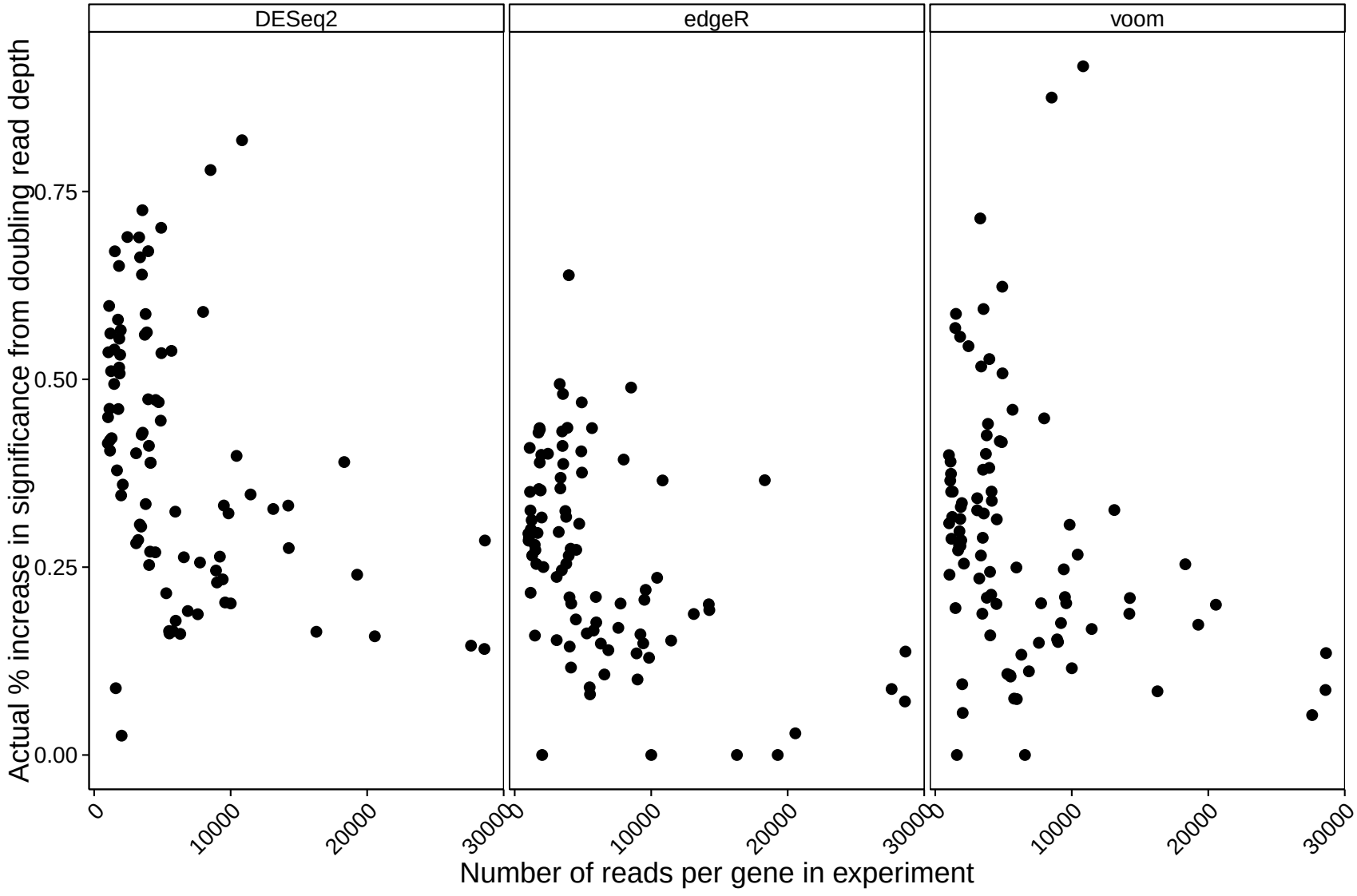
Results across all experiments



12 Expression Atlas experiments



Minimum read depth?



Conclusion

- We have a developed a model can help predict the increase in power from an increasing read depth
- Works well across 78 RNA-Seq read experiments that show varying degrees of saturations.
- A reasonable question: “Is there a minimum read depth that can be applied across all experiments?”

Acknowledgements

David Robinson
(QCB PhD student)

John D. Storey
(Principal Investigator)