Supplementary Methods

1 Chernoff bound to define prevalence at the sequencing depth employed by this study

Assume that event e occurs with probability p and further assume that there are n trials. Then, the expected number of occurrences of e is:

$$\mu = \mathbb{E}\left[\sum_{i=1}^{n} p\right] = np. \tag{1}$$

Given that the n trials are independent, Chernoff bound [1] applied to our setting tells us the following:

Theorem 1 *Given n independent trials, where in each trial event e occurs with probability p, then for $0 < \epsilon \leq 1$, the the probability of seeing X occurrences of e as a function of ϵ and μ (given in Equation (1)) is bounded as follows:*

$$Pr[X < (1-\epsilon)\mu] < e^{-\mu\epsilon^2/2}. \tag{2}$$

Given that we need to evaluate the probability of obtaining at least one sample of e with 95% confidence, we are interested in computing the value of the parameter μ (and thus also p) for which $\Pr[X < 1] < 0.05$. Thus, setting in Equation (2), $(1-\epsilon)\mu = 1$ we get that

$$\epsilon = \frac{\mu - 1}{\mu}.$$

Also, since we want $\Pr[X < 1]$ to be bounded by 0.05, we get that

$$\begin{aligned} e^{-\mu\epsilon^2/2} &\leq 0.05 & (3)\\ -\mu\epsilon^2/2 &\leq \log(0.05) & (4)\\ \mu^2 + (2\log(0.05) - 2)\mu + 1 &\geq 0. & (5) \end{aligned}$$

Solving the above inequality for μ (and after doing the necessary roundings) we get:

$$\mu \leq 0.1272$$

and

$$\mu \geq 7.86.$$

Clearly, for values of $\mu \leq 0.1272$, the value of $\epsilon = \frac{\mu-1}{\mu}$ is negative and thus these values of μ can be ignored. Thus, we conclude that as long as $\mu \geq 7.86$ we get at least one sample of e in our trials. If we want to compute the smallest value of p for which this is possible we get by Equation (1) that

$$pn \geq 7.86$$

or

$$p \geq \frac{7.86}{n}.$$

Plugging in $n = 1709$ we get that for $p > 0.0052$ we are 95% confident that we will see at least one sample of e.

References

[1] R. Motwani and P. Raghavan. *Randomized Algorithms.* Cambridge University Press, 1995.