

## Supplementary Methods for: A Scalable Gut Organoid Model Reveals the Genome-Wide Invasive Program of a Human Restricted Pathogen

Maria Letizia Di Martino<sup>1,\*</sup>, Laura Jenniches<sup>2,\*</sup>, Anjeela Bhetwal<sup>1</sup>, Jens Eriksson<sup>1</sup>, Ana C. C. Lopes<sup>1</sup>, Angelika Ntokaki<sup>1</sup>, Martina Pasqua<sup>1,3</sup>, Magnus Sundbom<sup>4</sup>, Martin Skogar<sup>4</sup>, Wilhelm Graf<sup>4</sup>, Dominic-Luc Webb<sup>5</sup>, Per M. Hellström<sup>5</sup>, André Mateus<sup>6,7</sup>, Lars Barquist<sup>2,8,9,\*</sup> and Mikael E. Sellin<sup>1,10,11,\*</sup>

<sup>1</sup> Department of Medical Biochemistry and Microbiology, Uppsala University, Sweden

<sup>2</sup> Helmholtz Institute for RNA-based Infection Research (HIRI), Helmholtz Centre for Infection Research (HZI), Würzburg, Germany

<sup>3</sup> Present address: Department of Biology and Biotechnologies "Charles Darwin", Sapienza University of Rome, Rome, Italy

<sup>4</sup> Department of Surgical Sciences, Uppsala University, Uppsala, Sweden

<sup>5</sup> Department of Medical Sciences, Uppsala University, Uppsala, Sweden

<sup>6</sup> Department of Chemistry, Umeå University, Sweden

<sup>7</sup> The Laboratory for Molecular Infection Medicine Sweden (MIMS), Umeå University, Sweden

<sup>8</sup> Faculty of Medicine, University of Würzburg, Würzburg, Germany

<sup>9</sup> Department of Biology, University of Toronto, Mississauga, ON, Canada

<sup>10</sup> Science for Life Laboratory, Uppsala, Sweden

<sup>11</sup> Lead contact

\*Correspondence: M.L.D.M. [ml.dimartino@imbim.uu.se](mailto:ml.dimartino@imbim.uu.se); L.J. [laura.jenniches@helmholtz-hiri.de](mailto:laura.jenniches@helmholtz-hiri.de); L.B. [lars.barquist@helmholtz-hiri.de](mailto:lars.barquist@helmholtz-hiri.de); M.E.S. [mikael.sellin@imbim.uu.se](mailto:mikael.sellin@imbim.uu.se)

### ZINB model for TraDIS screen

#### Definition of variables and parameters

$Y_{rgk,in/out}$ : input/output read count of insertion site  $k$  in gene  $g$  and replicate  $r$ .

$N_g$ : number of genes

$N_{k,r}$ : number of non-zero insertion sites in the input sample of replicate  $r$ .

$N_{r,in/out}$ : total number of read counts in the input/output of replicate  $r$ .

$N_r = \frac{N_{r,out}}{N_{r,in}}$ : ratio between sum of output and sum of input counts in replicate  $r$ .

$n_r$ : normalization factor between input and output sample of replicate  $r$ .

$\alpha_{rgk,in} = \log(Y_{rgk,in})$ : input log-counts, strictly positive, because insertion sites have been removed from the replicates where they have a zero-count in the input sample.

$\bar{\alpha}_{r,in} = \frac{1}{N_{k,r}} \sum_{g,k} \alpha_{rgk,in}$ : mean of input log-counts in replicate  $r$ .

$\tilde{\alpha}_{rgk,in} = \alpha_{rgk,in} - \bar{\alpha}_{r,in}$ : log-count normalized by mean count of replicate  $r$ .

$\alpha_{r,max} = \max(\tilde{\alpha}_{rgk,in})$ : maximum normalized log-count of replicate  $r$ .

**Model description.** We developed a Bayesian model to extract genome-wide gene-wise fitness scores in the presence of experimental bottlenecks (Fig 2D,S2F). The processed input counts align well with a negative binomial (NB) distribution (Fig S2H), a standard choice for modeling count data from next-generation sequencing technologies. However, the infection bottleneck introduces zero-inflation in the output data (Fig S2H), rendering the NB distribution inadequate for capturing the probability of observing a zero-count in the output libraries. Consequently, we employed a zero-inflated negative binomial (ZINB) model, constituting a mixture of an NB model and an additional technical “zero” component.

Given the unique mutant composition of individual replicates (Fig 2D, S2G), we encountered challenges in modeling gene-wise mean counts. Instead, we assumed that the output counts for a

single insertion site  $Y_{rgk,out}$  follow a ZINB distribution, where their probability of a stochastic (technical) zero probabilities due to experimental bottlenecks are absorbed into the given by the mixing coefficient  $\theta_{rgk}$ . The mean of the NB distribution depends on the normalized read count of the insertion site in the corresponding input sample and the gene-wise log-fold change  $\log FC_g$  between input and output abundance of all insertion sites in gene  $g$  across all replicates:

$$ZINB(Y_{rgk,out}; \mu_{rgk}, \phi_{rgk}) = \begin{cases} \theta_{rgk} + (1 - \theta_{rgk}) NB(0; 0, \phi_{rgk}), & \text{if } Y_{rgk,out} = 0 \\ (1 - \theta_{rgk}) NB(Y_{rgk,out}; \mu_{rgk}, \phi_{rgk}), & \text{if } Y_{rgk,out} > 0 \end{cases} \quad (1)$$

where  $\mu_{rgk} = \exp(\alpha_{rgk,in} + n_r + \log FC_g)$  denotes the expected output count. The model parameter  $\log FC_g$  uses the natural logarithm.

The technical zero probability  $\theta_{rgk}$  for the insertion site  $k$  in gene  $g$  and replicate  $r$  depends on its log-count in the input  $\alpha_{rgk,in}$  (Fig S2J) and the fraction of zero counts in the output library compared to the input library  $f_{z,r}$  (Fig S2I) and the log-count in the input  $\alpha_{rgk,in}$  (Fig S2J). We assumed the abundance-dependence to be linear and utilized a logit linker function, i.e.

$$\theta_{rgk} = \text{logit}^{-1}(f_{z,r} + a_1 + a_2 \cdot \alpha_{rgk,in}),$$

where  $a_1$  and  $a_2$  are model parameters.

Furthermore, we assumed that highly abundant insertion sites are measured more accurately, which we account for by making the dispersion coefficient of the NB model gene- and abundance-dependent

$$\phi_{rgk} = \phi_g \cdot \left(1 + b \cdot \frac{\tilde{\alpha}_{rgk,in}}{\alpha_{r,max}}\right),$$

where  $b \in [-1, 1]$  to ensure that the gene-wise dispersion is multiplied by a factor between 0 and 1. By integrating the technical zero probability into the mixing coefficient, zeros due to a fitness defect contributes to the gene-wise log-fold change, which represent the differential abundance in the output samples corrected for technical loss.

**Normalization and modeling of infection bottlenecks.** To correct for sequencing depth and infection bottlenecks, we extracted the normalization factors  $n_r$  and the mixing coefficients  $\theta_{rgk}$  from the 12,573 insertion sites located within pseudogenes. Given that pseudogenes are non-functional, mutations in these regions should not confer fitness advantages or defects, i.e.

$$\log FC_g = 0.$$

With a minimum of 145 insertion sites in pseudogenes per replicate, pseudogenes provide a sufficient basis for determining the normalization factor across all replicates.

As insertion sites are unique to each input-output pair, pairwise normalization factors  $n_r$  suffice. We corrected the normalization factor for differences in library size between input and output

$$n_r = \log(N_r) + \tilde{n}_r,$$

where  $\tilde{n}_r$  is a model parameter, representing a global log-fold change between input and output counts of replicate  $r$ .

Furthermore, this step involves capturing the dependence of the mixing coefficient  $\theta_{rgk}$  on the log-abundance of the insertion site in the input sample  $\alpha_{rgk,in}$  and the fraction of insertion sites with zero-count in the output  $f_{z,r}$ , as well as the abundance dependence of the dispersion  $\phi_{rgk}$ . Thus, we fit the parameters  $a_1$ ,  $a_2$  and  $b$  as defined above.

To improve the variance and normalization modeling by sharing information across genetic loci and sequencing libraries, we imposed a hierarchical structure:

$$\phi_g \sim N(\mu_\phi, \sigma_\phi) \text{ and } \tilde{n}_r \sim N(\mu_n, \sigma_n).$$

The remaining model parameters were assigned either Gaussian (N) or Cauchy (C) distributions, depending on the estimability of the parameters' magnitude prior to model fitting

$$\mu_\phi \sim C(1, 0.5), \sigma_\phi, \mu_n, \sigma_n \sim C(0, 1) \text{ and } a, b \sim N(0, 1).$$



Additionally, we enforce  $\mu_\phi, \sigma_\phi, \sigma_n \geq 0$  and  $b \in [-1, 1]$ , to ensure positive scale parameters.

**Extraction of gene-wise fitness scores.** After fixing the normalization factors and determining the abundance and replicate dependence of the mixing coefficient  $\theta_{rgk}$  within the ZINB model, we proceeded to fit the ZINB model (Eq 1) to the complete data set to extract the gene-wise log-fold changes  $\log FC_g$  and the gene-wise dispersion  $\phi_g$ . Again, we improved the variance modeling by imposing a hierarchical gaussian prior:

$$\phi_g \sim N(\mu_\phi, \sigma_\phi).$$

We set the scale parameter in the prior distribution of log-fold changes to a large value

$$\log FC_g \sim N(\mu_{\log FC}, 5)$$

to prevent undue shrinkage of the effect sizes. Importantly, we chose not to center the prior distribution at zero. In TraDIS screens, it is common for the number of disadvantageous mutations to exceed advantageous ones, leading to a negative mean  $\mu_{\log FC}$ . The hyper parameters followed Cauchy distributions, i.e.  $\mu_{\log FC}, \sigma_\phi \sim C(0, 1)$  and  $\mu_\phi \sim C(1, 0.5)$ . Again, we ensured positive values for the parameters  $\mu_\phi$  and  $\sigma_\phi$ . For figures and tables, we converted the model parameters  $\log FC_g$  to fitness scores ( $\log_2 FC_g$ ) changing from the natural logarithm to base 2.

**Fitting the Bayesian models to data from the TraDIS screen.** To derive the posterior distributions of the model parameters, we employed the probabilistic programming language Stan (v.2.31.0). The statistical models were fitted to the TraDIS screen data running two chains of 1000 Markov Chain Monte Carlo (MCMC) samples each (method=sample num\_samples=1000 num\_warmup=1000 adapt\_delta=0.95 algorithm=hmc engine=nuts max\_depth=12).

**Determine statistical significance of log-fold changes.** Fitting the Bayesian ZINB model with log-fold changes to the TraDIS screen data yields posterior distributions for the gene-wise log-fold changes. The expected gene-wise log-fold change is given by the median. To assess statistical significance and control the false discovery rate (FDR), we computed z-values

$$z_g = \frac{\log FC_g}{\Delta \log FC_g}, \text{ where } \Delta \log FC_g \text{ is the standard deviation of the posterior distribution. These z-values}$$

are expected to follow a normal distribution under the null hypothesis (<http://dx.doi.org/10.1214/07-STS236>). Indeed, the distribution of the z-values (Fig S2KJ) closely resembles a standard normal distribution with an excess of negative z-values which correspond to the genes of interest, i.e. potential invasion factors.

Assuming that at least 10% of genes play a role during infection (a rather conservative assumption), the FDR for a negative z-value cutoff is determined by

$$\text{FDR}(z) = 2 \cdot \frac{0.9 \cdot \Phi(z|0,1)}{\left| \{z_g | z_g \leq z\} \right|},$$

where  $\Phi$  is the cumulative normal distribution and  $\left| \{z_g | z_g \leq z\} \right|$  the number of genes with  $z_g \leq z$ . Similarly, the FDR for positive z-values is given by

$$\text{FDR}(z) = 2 \cdot \frac{0.9 \cdot \Phi(-z|0,1)}{\left| \left\{ z_g \mid z_g \geq z \right\} \right|}.$$

The resulting curve is depicted in Fig S2L [K](#).

We explored two alternative methods to assign FDR to the log-fold changes each presenting its own challenges. One approach involves calculating Bayesian p values, representing the fraction of the 2000 MCMC samples  $s$  in agreement with the null hypothesis. However, due to the precision limitations inherent in measuring log-fold changes in the experimental setup, the finite interval  $[-h_0, h_0]$  is assumed to align with the null hypothesis, resulting in the following Bayesian p values

$$p_g = \begin{cases} \frac{\left| \left\{ s \mid s \geq -h_0 \right\} \right|}{2000}, & \text{if } \log\text{FC}_g < 0 \\ \frac{\left| \left\{ s \mid s \leq h_0 \right\} \right|}{2000}, & \text{if } \log\text{FC}_g > 0 \end{cases}.$$

This approach poses the challenge of defining the distribution of p values under the null hypothesis, a prerequisite for converting Bayesian p values into FDRs.

Alternatively, a more elegant method involves the use of Bayes' factors (Jeffreys, H (1961). Theory of probability, 3rd Edn. Oxford: Oxford University Press.). Bayes' factors quantify how much less likely the null hypothesis is compared to the median log-fold change after fitting the model to the data as compared to the prior estimates

$$\text{Bayes' factor}(\log\text{FC}_g) = \frac{\Pr(\text{data}|0)}{\Pr(\text{data}|\log\text{FC}_g)} = \frac{\Pr(0|\text{data})\Pr(\log\text{FC}_g)}{\Pr(\log\text{FC}_g|\text{data})\Pr(0)}.$$

While a Bayes' factor smaller than 0.1 is considered strong evidence, this is not inherently corrected for multiple hypothesis testing. Considering the challenge associated with defining the null hypothesis of Bayesian p values and correcting Bayes' factors for multiple hypothesis testing, we opted to use z-values to assign FDRs to log-fold changes.