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

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

[1] *Department of Medical Biochemistry and Microbiology, Uppsala University, Sweden*
[2] *Helmholtz Institute for RNA-based Infection Research (HIRI), Helmholtz Centre for Infection Research (HZI), Würzburg, Germany*
[3] *Present address: Department of Biology and Biotechnologies "Charles Darwin", Sapienza University of Rome, Rome, Italy*
[4] *Department of Surgical Sciences, Uppsala University, Uppsala, Sweden*
[5] *Department of Medical Sciences, Uppsala University, Uppsala, Sweden*
[6] *Department of Chemistry, Umeå University, Sweden*
[7] *The Laboratory for Molecular Infection Medicine Sweden (MIMS), Umeå University, Sweden*
[8] *Faculty of Medicine, University of Würzburg, Würzburg, Germany*
[9] *Department of Biology, University of Toronto, Mississauga, ON, Canada*
[10] *Science for Life Laboratory, Uppsala, Sweden*
[11] *Lead contact*
*Correspondence: M.L.D.M. ml.dimartino@imbim.uu.se; L.J. laura.jenniches@helmholtz-hiri.de; L.B. lars.barquist@helmholtz-hiri.de; M.E.S. mikael.sellin@imbim.uu.se

ZINB model for TraDIS screen

Definition of variables and parameters

$Y_{rgk,\text{in/out}}$: 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.

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

$N_r = \frac{N_{r,\text{out}}}{N_{r,\text{in}}}$: 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.

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

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

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

$\alpha_{r,\max} = \max\left(\tilde{\alpha}_{rgk,\text{in}}\right)$: 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 $Y_{rgk,\text{out}}$ follow a ZINB distribution, where their probability of a stochastic (tech-

nical) zero due to experimental bottlenecks is given by the mixing coefficient θ_{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 $\log\mathrm{FC}_g$ between input and output abundance of all insertion sites in gene g across all replicates:

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

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

The technical zero probability θ_{rgk} for the insertion site k in gene g and replicate r depends on 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 $\alpha_{rgk,\mathrm{in}}$ (Fig S2J). We assumed the abundance-dependence to be linear and utilized a logit linker function, i.e.

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

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

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

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 θ_{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.

$$\log\mathrm{FC}_g=0.$$

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

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

where $\tilde{\mathrm{n}}_r$ 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 θ_{rgk} on the log-abundance of the insertion site in the input sample $\alpha_{rgk,\mathrm{in}}$ and the fraction of insertion sites with zero-count in the output $f_{z,r}$, as well as the abundance dependence of the dispersion ϕ_{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:

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

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
$\mu_\phi \sim C(1,0.5)$, $\sigma_\phi, \mu_n, \sigma_n \sim C(0,1)$ 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 θ_{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 $\mathrm{logFC}_g$and the gene-wise dispersion ϕ_g. Again, we improved the variance modeling by imposing a hierarchical gaussian prior:

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

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

$$\mathrm{logFC}_g \sim N\left(\mu_{\mathrm{logFC}}, 5\right)$$

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 μ_{logFC}. The hyper parameters followed Cauchy distributions, i.e. $\mu_{\mathrm{logFC}}, \sigma_\phi \sim C(0,1)$ and $\mu_\phi \sim C(1,0.5)$. Again, we ensured positive values for the parameters μ_ϕ and σ_ϕ. For figures and tables, we converted the model parameters logFC_g to fitness scores ($\log_2 \mathrm{FC}_g$) 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 $z_g = \frac{\mathrm{logFC}_g}{\Delta \mathrm{logFC}_g}$, where $\Delta \mathrm{logFC}_g$ 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 S2K) 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

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

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

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

The resulting curve is depicted in Fig S2L.

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

$$p_g = \begin{cases} \frac{\left|\left\{s \middle| s \geq -h_0\right\}\right|}{2000}, \text{ if } \mathrm{logFC}_g < 0 \\ \frac{\left|\left\{s \middle| s \leq h_0\right\}\right|}{2000}, \text{ if } \mathrm{logFC}_g > 0 \end{cases}.$$

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

$$\text{Bayes' factor}\left(\mathrm{logFC}_g\right) = \frac{\Pr(\text{data}|0)}{\Pr\left(\text{data}\middle|\mathrm{logFC}_g\right)} = \frac{\Pr(0|\text{data})\Pr\left(\mathrm{logFC}_g\right)}{\Pr\left(\mathrm{logFC}_g|\text{data}\right)\Pr(0)}.$$

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.