nature microbiology

SUPPLEMENTARY INFORMATION

VOLUME: 2 | ARTICLE NUMBER: 17088

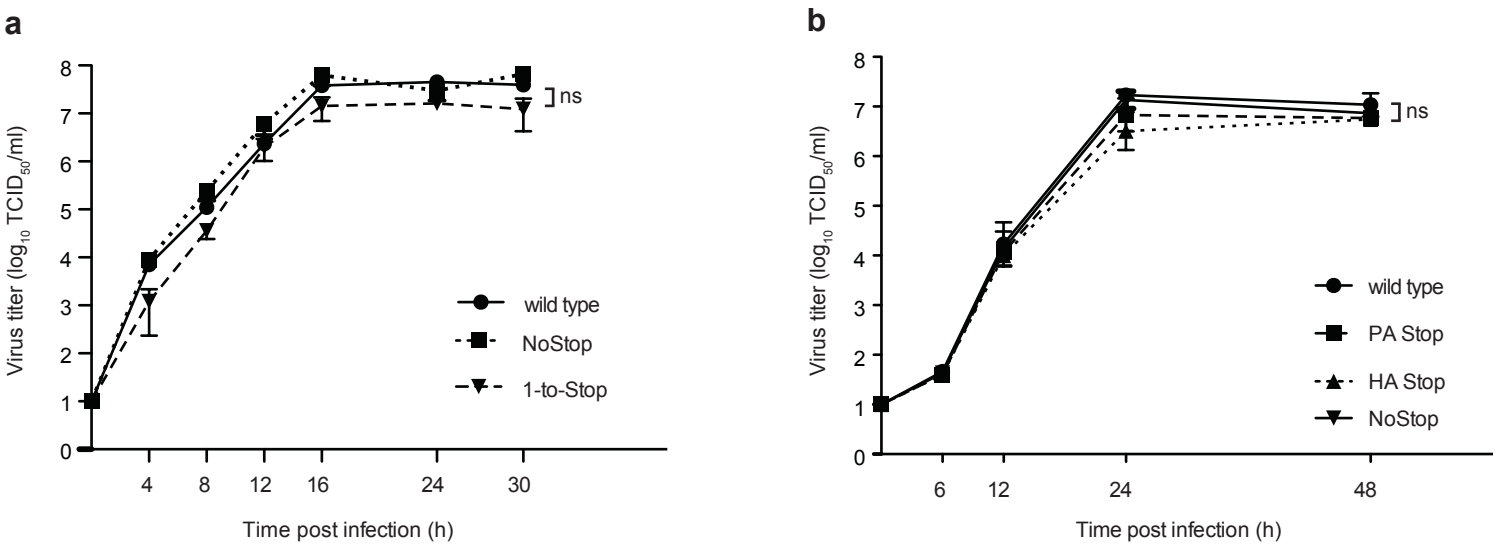
In the format provided by the authors and unedited.

Attenuation of RNA viruses by redirecting their evolution in sequence space

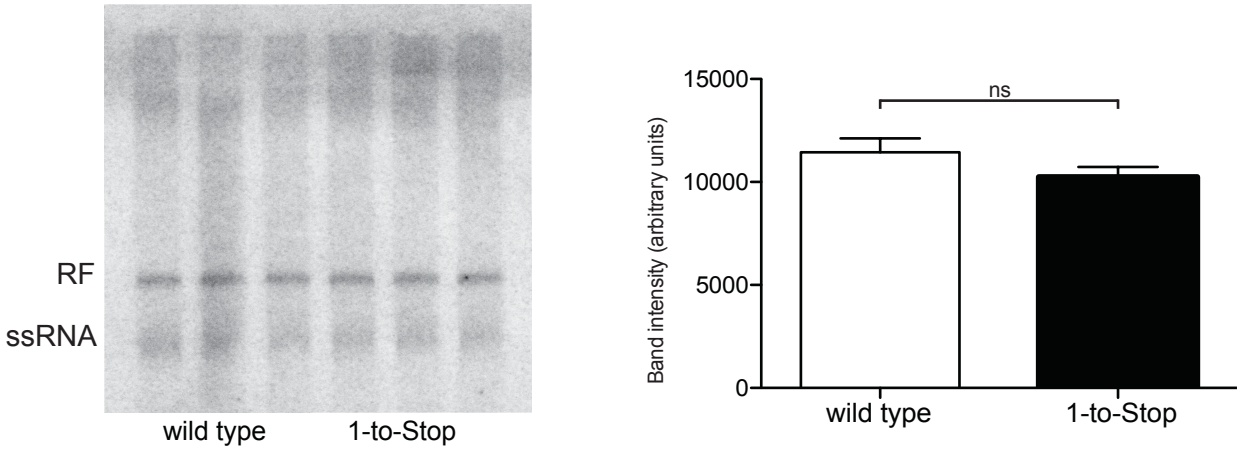
Gonzalo Moratorio, Rasmus Henningsson, Cyril Barbezange, Lucia Carrau, Antonio V. Bordería, Hervé Blanc, Stephanie Beaucourt, Enzo Z. Poirier, Thomas Vallet, Jeremy Boussier, Bryan C. Mounce, Magnus Fontes and Marco Vignuzzi

Supplementary Information Guide

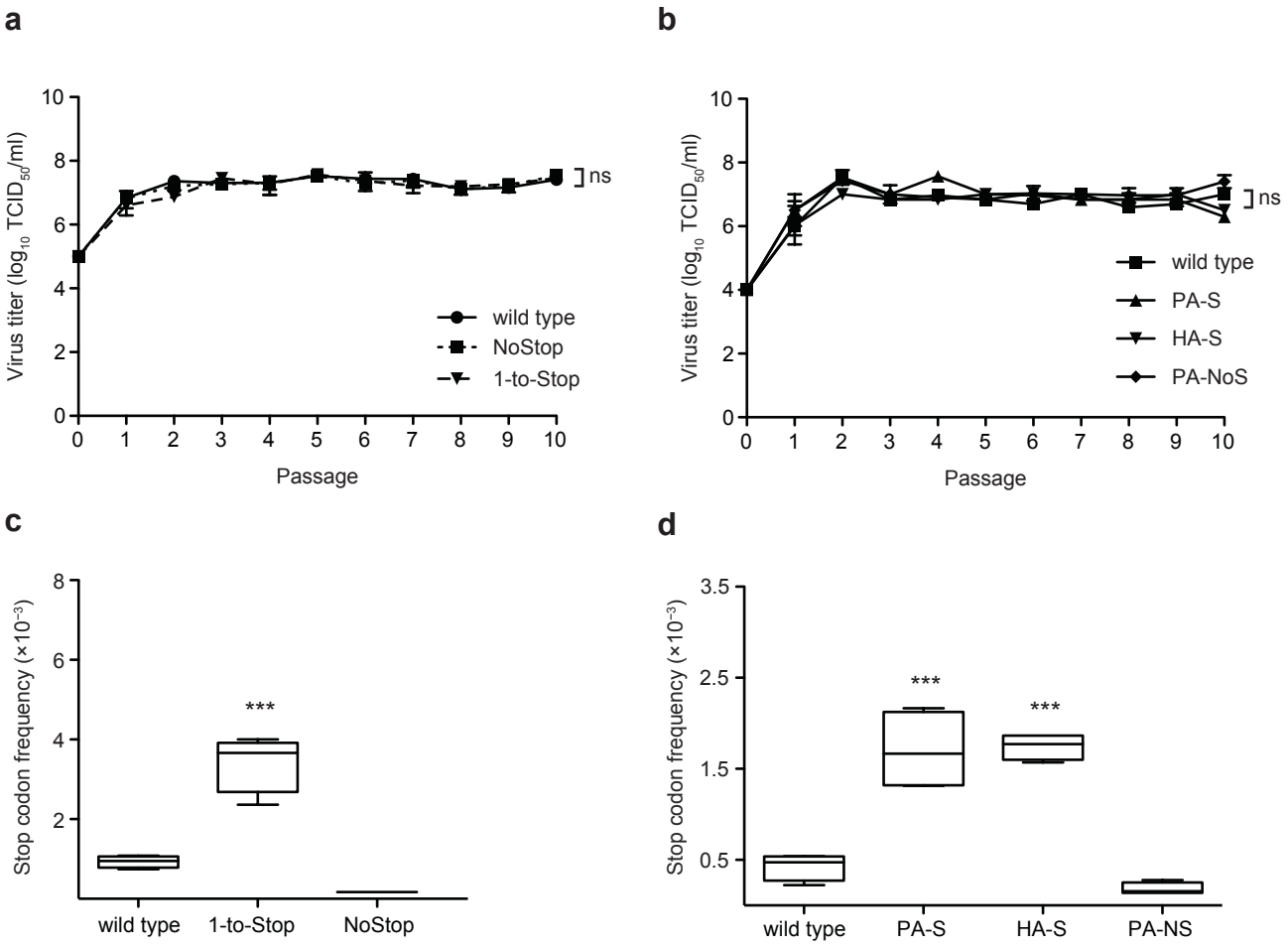
Supplementary Table 1 – provided as a separate Excel document



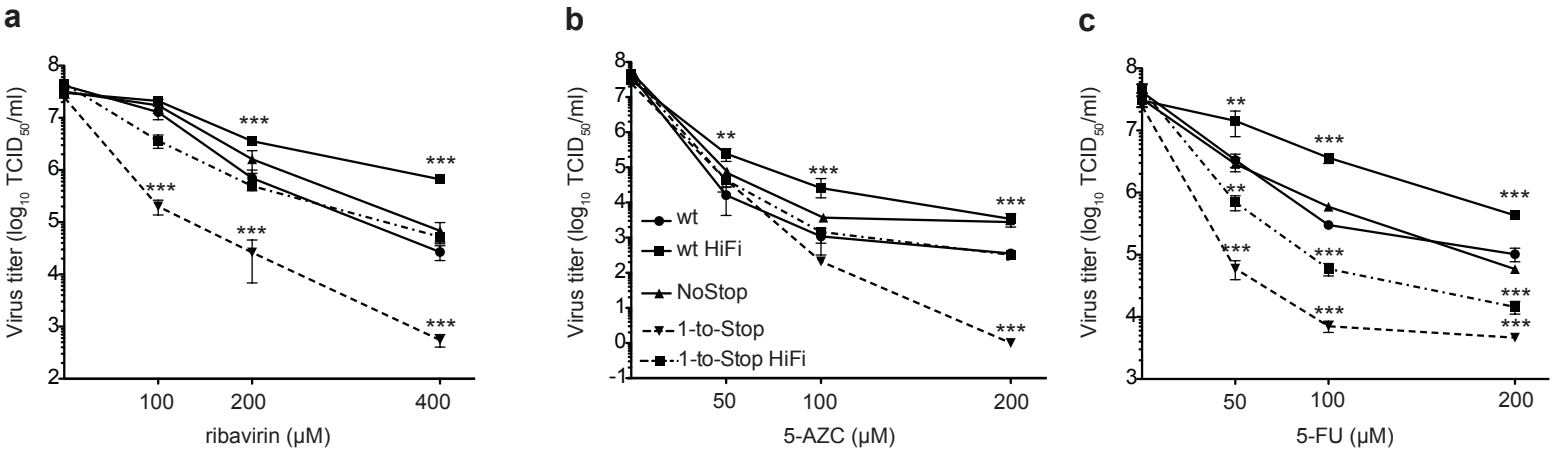
Supplementary Figure 1. (**a**) Replication kinetics of wild type, 1-to-Stop and NoStop Coxsackie virus B3 in HeLa cells infected at MOI 0.1. (**b**) Replication kinetics of wild type, 1-to-Stop PA and HA, and NoStop PA influenza A viruses in MDCK cells infected at MOI 0.1. Bars show mean and SEM; *n* = 3 per group. ns, non-significant (two-tailed unpaired t-test with Bonferroni correction comparing wild type to each mutant).



Supplementary Figure 2. *In vitro* RNA replication assay. Membranes containing replication complexes from HeLa cells infected with wild type or 1-to-Stop CVB3 viruses were purified and used for *in vitro* replication assays, by adding *in vitro* transcribed RNA, corresponding to wild type (3 samples) or 1-to Stop (3 samples) CVB3 viruses, and radiolabeled UTP. The Replicative Form (RF) and single stranded RNA (ssRNA) are visualized. Density of each band was determined by ImageJ. Bars show mean and SEM; $n = 3$ per group. ns, non-significant (two-tailed unpaired t-test).



Supplementary Figure 3. Genetic and phenotypic stability of 1-to-Stop and NoStop viruses after serial passage in tissue culture. (**a**, **b**) Virus titres over 10 serial passages. HeLa (**a**) or MDCK (**b**) cells were infected with CVB3 (**a**) or influenza A (**b**) variants at MOI 0.1. Virus titres were determined for each passage. Mean and SEM are shown; $n = 3$. ns, non-significant (two-way analysis of variance) (**c**,**d**) Frequency of Stop mutations observed in deep sequencing reads from wild type, 1-to-Stop, and NoStop variants from passage number 10 for CVB3 variants (**c**) or influenza A variants (**d**). Boxes show median and interquartile range, whiskers range or 1.5 interquartile range in case of outlier, individual dots indicate outliers; $n = 6$ per group. ***$P < 0.001$ (two-tailed unpaired t-test with Bonferroni correction, comparing wild type to each mutant).



Supplementary Figure 4. 1-to-Stop high-fidelity CVB3 recovers wild type phenotype in presence of mutagens.

Sensitivity of wild type (wt), wild type high-fidelity (wt HiFi), NoStop, 1-to-Stop and 1-to-Stop HiFi CVB3 viruses to increasing concentrations of (**a**) ribavirin, (**b**) 5-azacytidine (5-AZC) and (**c**) 5-fluorouracil (5-FU). Graphs show mean and SEM; $n = 3$ per group. $^{**}P < 0.01$, $^{***}P < 0.001$ (two-way analysis of variance with Bonferroni post test).

Supplementary Methods

Mathematical assessment of stop codon background noise

The overall NGS error frequency in the context of our work, is not of interest *per se*; rather, only the fraction of errors that cause stop codons, since only those errors could affect our results. The question we ask here is whether more of fewer Stop mutations are observed at all of the altered Ser/Leu sites combined, rather than at individual sites. This increases our sample size by approximately 100-fold, thereby further increasing our statistical power. To model the frequency of 'false' stop codons due to sequencing error, we use a Poisson background noise model (a standard for estimating independent rare errors under minimal additional structural assumptions, which has already been used to model NGS error rates in the context of minority variant discovery)[1,2]. Since sequencing errors are nucleotide-context dependent[3], and the number of stop codons that can be reached by a single mutation is different for different 1-to-stop codons, we model y_{ij}, the number of observed stop codons for sample i at site j, as an observation of the random variable:

$$Y_{ij} \sim \mathrm{Po}\left(N_{ij}(\lambda_C + \mu_{ij})\right),$$

where N_{ij} is the total number of reads for the same sample and site, λ_C is the probability of observing an erroneous stop codon given the nucleotide context C, μ_{ij} is the true stop codon frequency for sample i at site j; and the sequencing errors in different reads are assumed to be independent. The Poisson approximation is very good since N_{ij} is large and $\lambda_C + \mu_{ij}$ is small. Both the influence of nucleotide context and the number of different stop codons reachable within one mutation are captured in the λ_C parameter.

Let $\nu_{ij} := N_{ij}(\lambda_C + \mu_{ij})$ and $P_\nu(x)$, the probability mass function of the Poisson distribution with rate ν. The log-likelihood is

$$l(\boldsymbol{\nu}; \boldsymbol{y}) = \sum_{i,j} \log P_{\nu_{ij}}(y_{ij}).$$

Hence, maximizing l is equivalent to maximizing each term separately, and the ML estimate for the Poisson distribution is given by $\hat{\nu}_{ij} = y_{ij}$. Reparameterizing in λ_C and $\boldsymbol{\mu}$, it follows that the ML estimate is achieved for any $\hat{\lambda}_C \in [0, \min_{i,j} y_{ij}/N_{ij}]$, with $\hat{\mu}_{ij} = y_{ij}/N_{ij} - \hat{\lambda}_C$, since $\lambda_C \geq 0$ and $\mu_{ij} \geq 0$ for all i, j. The log likelihood test statistic

$$D(\lambda_C^0) = 2\left(\max_{\lambda_C, \boldsymbol{\mu}} l(\lambda_C, \boldsymbol{\mu}) - \max_{\boldsymbol{\mu}} l(\lambda_C^0, \boldsymbol{\mu})\right),$$

measures the drop in log likelihood between the full model and a reduced model with λ_C fixed at λ_C^0. By profile likelihood, a 95% confidence interval for λ_C consists of all λ_C^0 such that $D(\lambda_C^0) \leq \chi_1^2(0.95)$, where χ_1^2 is the quantile function for the χ^2 distribution with one degree of freedom, since the change in model order is one. Now, since $D(0) = 0$ and D is a decreasing function, the upper endpoint of the profile likelihood confidence interval for λ_C can be found by a binary search. Note how the ML estimate of each μ_{ij} in the reduced model is still given by $y_{ij}/N_{ij} - \hat{\lambda}_C$, but constrained such that $\mu_{ij} \geq 0$.

The model was thus applied to 5 (96-well plates each) sequencing runs, for a total of 420 samples. Only samples from the same run are compared, to ignore noise due to batch effect. Each codon site that had a 1-to-Stop codon in the wild type genome (and thus the identical codon, in the identical context, in the 1-to-Stop construct) was considered. The nucleotide contexts coincide with the 1-to-Stop codons (UUA, UUG, UCA and UCG), since a mutation that produces a stop codon must change the middle nucleotide in the codon. To produce a single confidence interval for the stop codon error rate λ for each sequencing run, the confidence intervals for the different nucleotide contexts, λ_{UUA}, λ_{UUG}, λ_{UCA} and λ_{UCG}, were averaged with weights proportional to the number of codon sites for each context.

Sequencing Run	**Weighted 95% confidence interval for λ**
CVB3 in vitro	$[0, 1.062 \cdot 10^{-6}]$
CVB3 in vivo	$[0, 2.098 \cdot 10^{-6}]$
Flu HA in vitro	$[0, 2.237 \cdot 10^{-6}]$
Flu PA in vitro	$[0, 1.135 \cdot 10^{-7}]$
Flu HA in vivo	$[0, 5.631 \cdot 10^{-6}]$
Flu PA in vivo	$[0, 2.031 \cdot 10^{-7}]$

Thus, considering the 1-to-Stop codon sites, if the sequencing error rates were at the upper endpoints of the 95% confidence intervals, the contribution of sequencing errors to the stop codon frequency as it was computed (sum over all Leu/Ser) would be $\sim 10^{-6} \times 100 = 10^{-4}$ for the 1-to-Stop virus, which is still more than 10 times below the observed frequencies.

Supplementary References

1. Ord, J. K. & Haight, F. A. Handbook of the Poisson Distribution. *OR* **18,** 478 (1967).
2. Raymond, S. *et al.* Performance comparison of next-generation sequencing platforms for determining HIV-1 coreceptor use. *Scientific Reports* **7,** 42215 (2017).
3. Welkers, M. R. A., Jonges, M., Jeeninga, R. E., Koopmans, M. P. G. & de Jong, M. D. Improved detection of artifactual viral minority variants in high-throughput sequencing data. *Front Microbiol* **5,** 804 (2014).