A Appendix

A.1 Assessing spatial heterogeneity

A typical commercial broiler chicken house in Pennsylvania is 400 ft × 50 ft. An important consideration during sample collection is therefore whether virus concentrations are spatially variable. We explored this possibility through an observational study. We divided houses into six arbitrary sections as shown in fig. S1. In each section, 5 dust samples were collected from wooden ledges located along the wall of the house, and in sections that also contained fans, we collected an additional five samples from the louvers that cover the fans. This collection method was performed in 2 houses, for 2 flocks of birds. In each case, three sections contained wooden ledges in addition to fans, and three sections contained wooden ledges but no fans. Virus quantities were measured using the qPCR method described in the main text. Here we used cycle threshold scores (CT scores) as our estimate of virus concentration. CT scores are inversely correlated with the log of virus concentration, such that small CT scores represented high concentrations of virus. Values of 40 were assigned to samples that never crossed the fluorescence threshold, because our qPCR assay was run for a total of 40 cycles.

We tested for heteroscedasticity, that is a change in variance with the mean, in the CT scores by examining the mean squared difference in CT score between our replicate samples against the mean in CT score for the two samples. A likelihood ratio test found a significant effect of mean CT score on the mean squared difference ($\chi^2 = 5.18, d.f. = 1, p = 0.023$, fig. S2). Nonetheless the effect was small relative to the variation in our data ($R^2 = 0.03$). We therefore ignored this heteroscedasticity when testing for effects of "location" and "substrate" in our data.

The data were divided into three balanced datasets for analysis. The first dataset contained all samples collected from wooden ledges, and no fan samples, hereafter referred to as

"wood data". The second dataset contained all samples collected from fans, and no wooden ledge samples, hereafter referred to as "fan data". The third dataset contained all samples collected from fans, and the samples that were collected from wooden ledges in sections that also contained fans, hereafter referred to as "fan and wood data".

We next constructed linear mixed effects models, and fit them in "R" using the function "lmer" in the package "lme4" (Bates et al., 2015). These models contained fixed effects of "House", "Flock", and the interaction between them. For the fan and wood data, we additionally included a fixed effect of "Substrate" to allow for differences in CT score between samples collected from wooden ledges and those collected from fan louvers. All models also contained a random effect of "Sample". For the wood data and fan data, we were interested in determining the significance of "Section", and so we also constructed models that also included or excluded this factor. Section was nested in house and flock. These simpler models were compared to their respective more complex models using likelihood ratio tests. We found that "Section" was highly significant in the wood data ($\chi^2 = 104.16, d.f. = 1, p < 0.001$), but it was not significant in the fan data ($\chi^2 = 1.76, d.f. = 1, p = 0.18$). This suggests that fan louver dust samples were more well mixed than wooden ledge dust samples, presumably, because the fans were designed to pull air from throughout the house. For the wood and fan data, we compared two models that contained all factors above, where one also included an effect of "Substrate". This analysis showed a significant effect of substrate ($\chi^2 = 43.2, d.f. = 1, p < 0.001$), where virus concentration was higher on samples collected from wood samples than from fan samples. These analyses suggest that dust samples collected from fan louvers would show less spatial variation than samples collected from wooden surfaces, and that virus concentrations would differ when collected from wooden ledges or fan louvers. We therefore used only samples collected from fan louvers during our surveillance. Despite this preliminary data suggesting that fans do not show spatial variation, we were conservative in our collection methods by sampling from the louvers of multiple fans

during each collection trip.

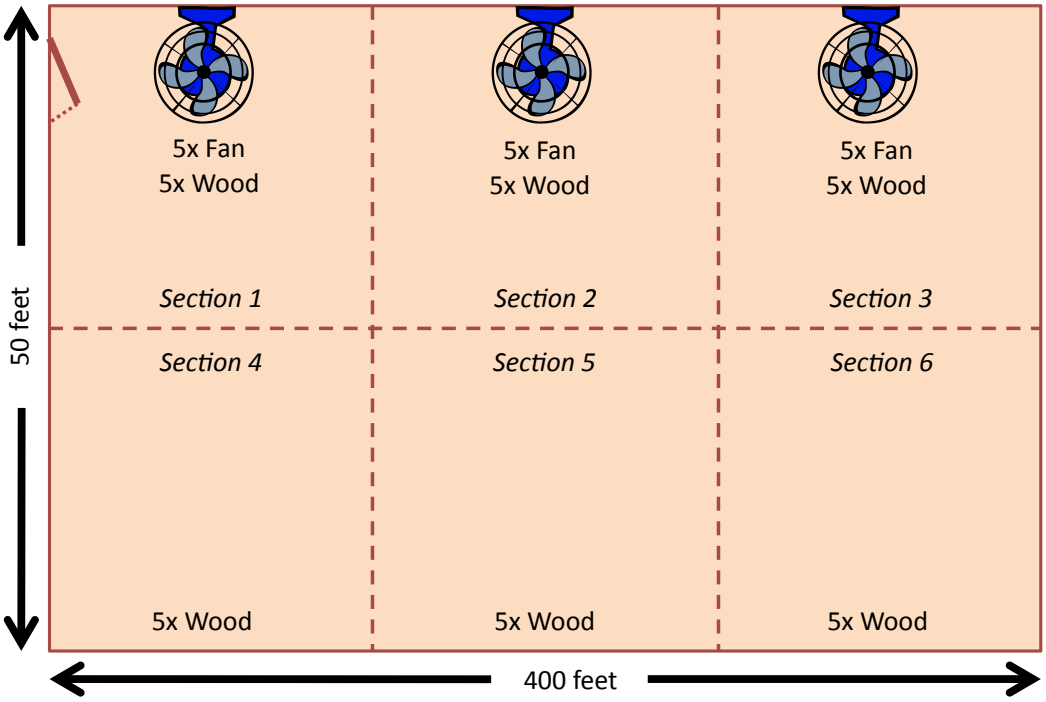

Figure S1: Schematic of a typical commercial broiler chicken house. Dotted lines illustrate partitioning used in the above sampling experiment, and do not represent physical separators within the houses.

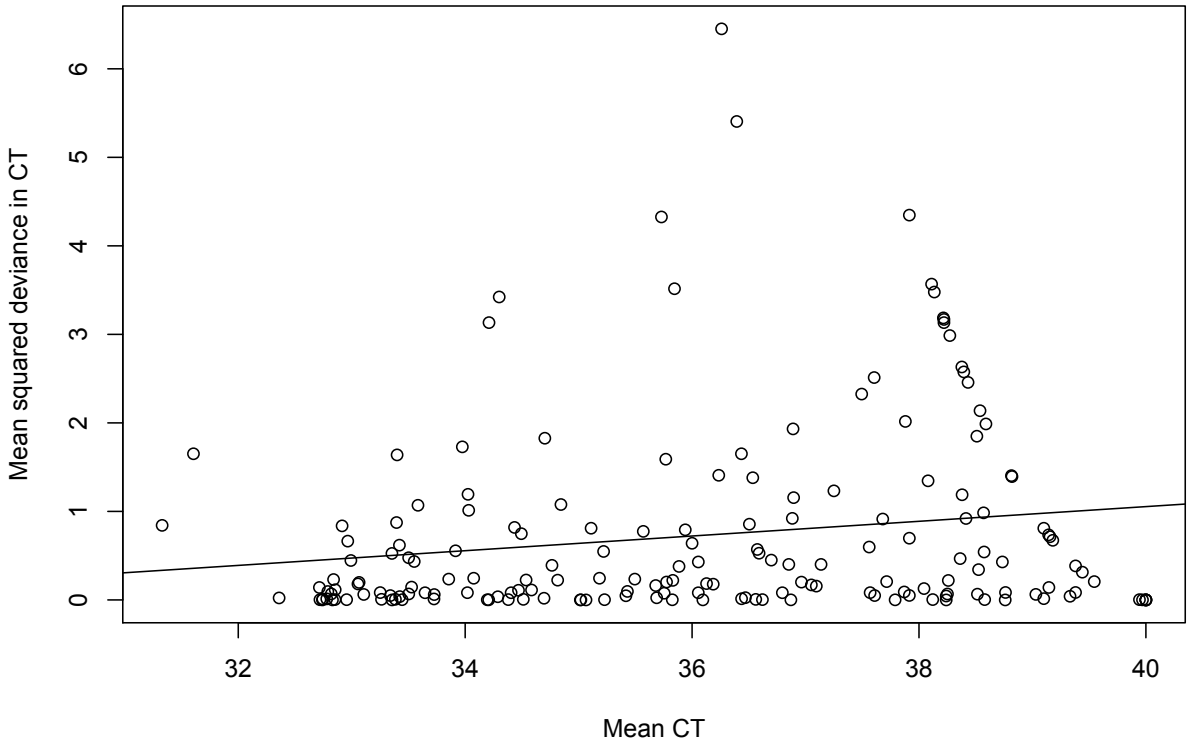

Figure S2: Mean square deviance of biological replicates as a function of CT score. The solid line is the best fit regression line.

A.2 Rispens virus interference

The Rispens vaccine is an attenuated strain of Marek's disease virus, and being a live virus, it is shed from vaccinated hosts (Islam et al., 2013; Rispens et al., 1972). In Pennsylvania, this vaccine is rarely used in broiler chicken flocks, but it is commonly used in broiler-breeder and egg-laying chickens. Due to its genetic similarity to wild-type virus, our qPCR assay was not able to perfectly discriminate between Rispens virus and wild-type virus. We assessed the degree of this interference through a lab experiment.

In this experiment we generated samples to test for interference. Each sample contained

one of eight concentrations of wild-type Marek's disease DNA from our virus standard ranging by tenfold dilutions from 2.41×10^7 to 2.41×10^0 DNA copies per 4 μl volume, or a control that lacked this DNA. The samples also contained one of eight concentrations of DNA from our Rispens virus standard ranging by tenfold dilutions from 1.60×10^7 to 1.60×10^0 DNA copies per 4 μl volume, or a control. By making every combination of the above, we generated 81 total samples. We then ran our wild-type virus assay and Rispens virus assay on each sample to quantify potential interference. The wild-type assay was run using the methods described in the main text. The Rispens virus assay was run in an identical manner, except that we used the Rispens virus DNA specific probe pp38-CVI from Baigent et al. (2016). We used the samples lacking either Rispens or wild-type Marek's disease virus DNA as standards for DNA quantification.

This analysis showed that in both the Rispens virus and wild-type virus assays interference was negligible when the target DNA was more prevalent than the non-target DNA (fig. S3). Quantification was fairly accurate, although counts were slightly biased low, when non-target DNA was approximately tenfold more prevalent than target DNA. Beyond this ratio, interference from the non-target DNA began to severely bias quantification. This bias was more extreme in the Rispens assay than in the wild-type assay.

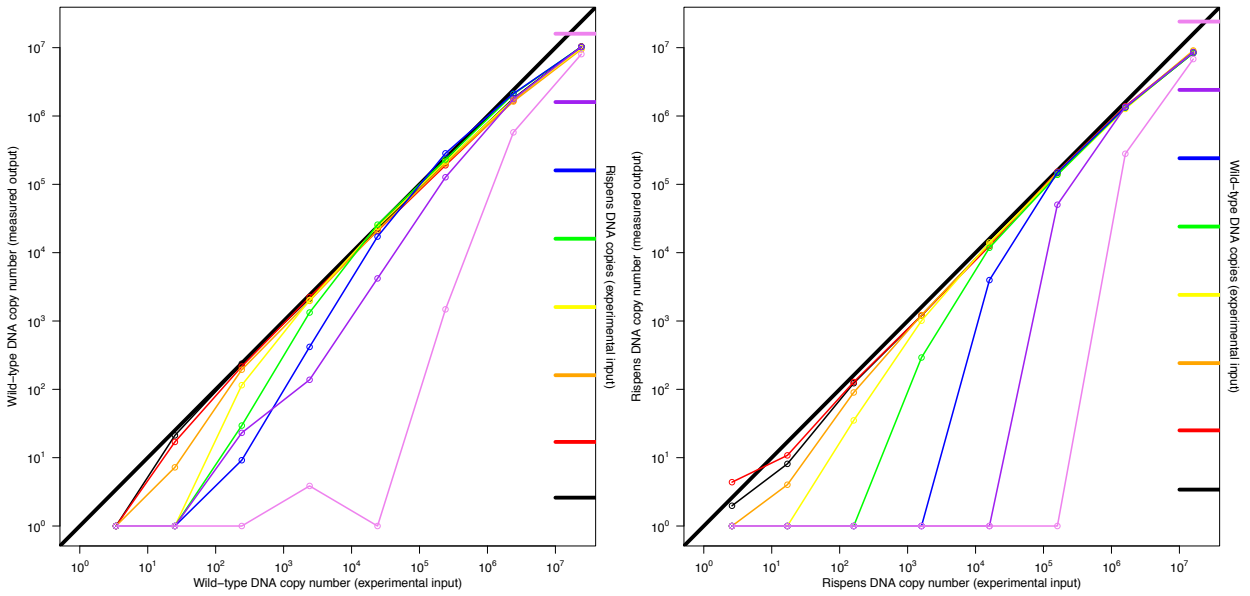

Figure S3: Interference plots for the wild-type virus assay (left), and Rispens virus assay (right). Each plot shows the quantification of target DNA over 7 orders of magnitude in the presence of contaminating wild-type or Rispens virus DNA. The diagonal black line is a 1:1 line representing perfect agreement between the assay and the experimental target value. The greater the deviance from this line, the less accurate the quantification. The data are shown as colored dots connected by same colored lines, where each color is a different level of interfering DNA. The quantity of interfering DNA is shown on the right side of the plot. Note that for both assays, quantification was fairly accurate when interfering DNA was less prevalent than target DNA, and only slightly biased when interfering DNA was 10-fold more prevalent.

A.3 Rispens virus in field samples

752 In Pennsylvania and many other parts of the world, the Rispens virus vaccine is used more commonly for breeder and layer birds than for broiler birds. As expected, we detected Rispens virus more often in samples collected from breeder and layer flocks than from broiler flocks (fig. S4). This, coupled with the observation that Rispens virus DNA can interfere with 756 detection of wild-type DNA in qPCR, suggested that observed differences in wild-type virus incidence between different production types might result from a detection bias. From our

model comparison, we concluded that wild-type virus was more prevalent in dust collected from broiler flocks than layer flocks. We thus performed a bootstrap analysis to show that this difference cannot be explained by a detection bias caused by qPCR interference.

In Appendix A.2 we showed that interference for both the Rispens and wild-type assays was negligible when target DNA was at least as prevalent as the contaminating DNA, and interference was minor when the non-target DNA was 10-fold greater than the target. We therefore wanted to determine whether the wild-type virus DNA detected in broiler samples would have reached or exceeded these ratios if they had a distribution of Rispens virus DNA similar to that seen in the layer samples.

To determine the distribution of Rispens virus concentration for our bootstrap analysis, we discarded all layer samples in which the wild-type virus concentration exceeded the Rispens virus concentration, and retained the measurements of Rispens virus concentration in all other layer samples. This left us with 311 measures of Rispens virus concentration. To determine the distribution of wild-type virus concentration, we retained the measurements of wild-type virus concentration from all broiler samples. This left us with 1304 measures of wild-type virus concentration. In our original analysis we had 317 measurements from layer samples, 2.5% of which had detectable wild-type virus. We drew 317 pairs of Rispens virus concentration and wild-type virus concentration with replacement and determined the fraction of these pairs that had "detectable wild-type virus". This process was repeated 100,000 times. Detectable wild-type virus can be interpreted to mean wild-type virus that is no more than 10-fold less prevalent than Rispens virus. However, to be conservative, we reran the analysis under the assumption that wild-type virus would only be detectable if it were more common than Rispens virus. In all 100,000 runs of both analyses, wild-type virus was detectable in a larger fraction of samples than that of the layers (fig. S5). This analysis showed that even with Rispens contamination, the prevalence of wild-type virus in layer samples was below that seen in broiler samples, suggesting that qPCR interference could not

784 explain the difference in virus detection between layer samples and broiler samples.

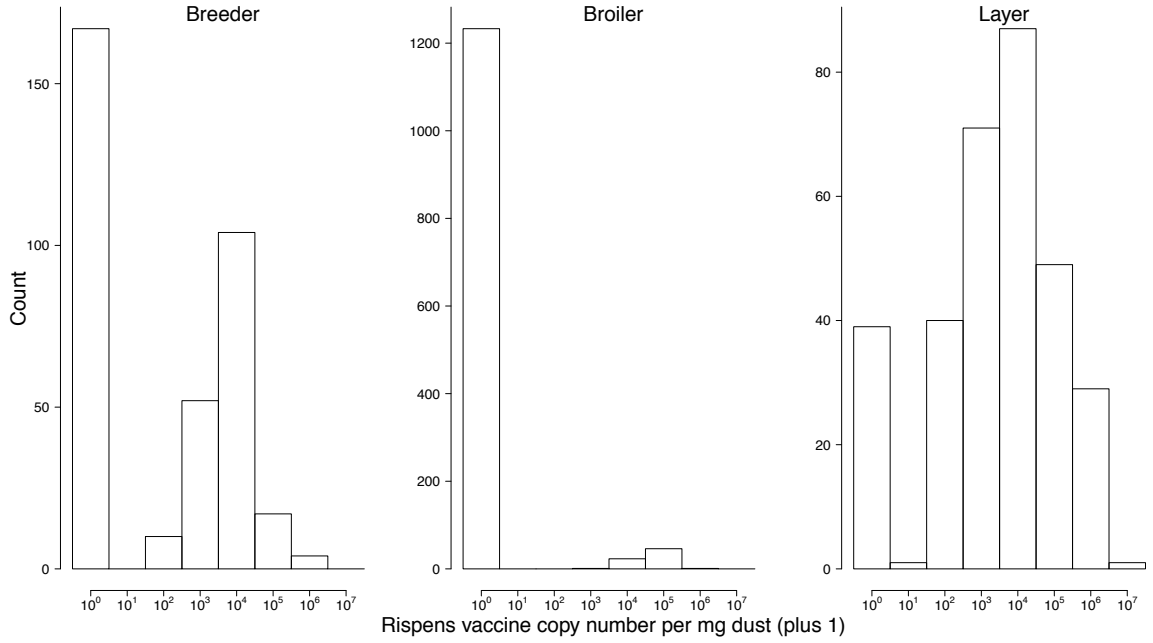

Figure S4: Measured concentration of Rispens virus DNA in samples. All breeder and layer samples, as well as broiler samples suspected to contain Rispens virus, were tested for the Rispens virus. Note, that for many of the broiler samples, qPCR curves were examined to determine whether Rispens virus was likely to be present at high concentrations. When no evidence of Rispens virus was present, these samples were assumed to have no Rispens virus present, because Rispens virus vaccine was not used in these chicken flocks.

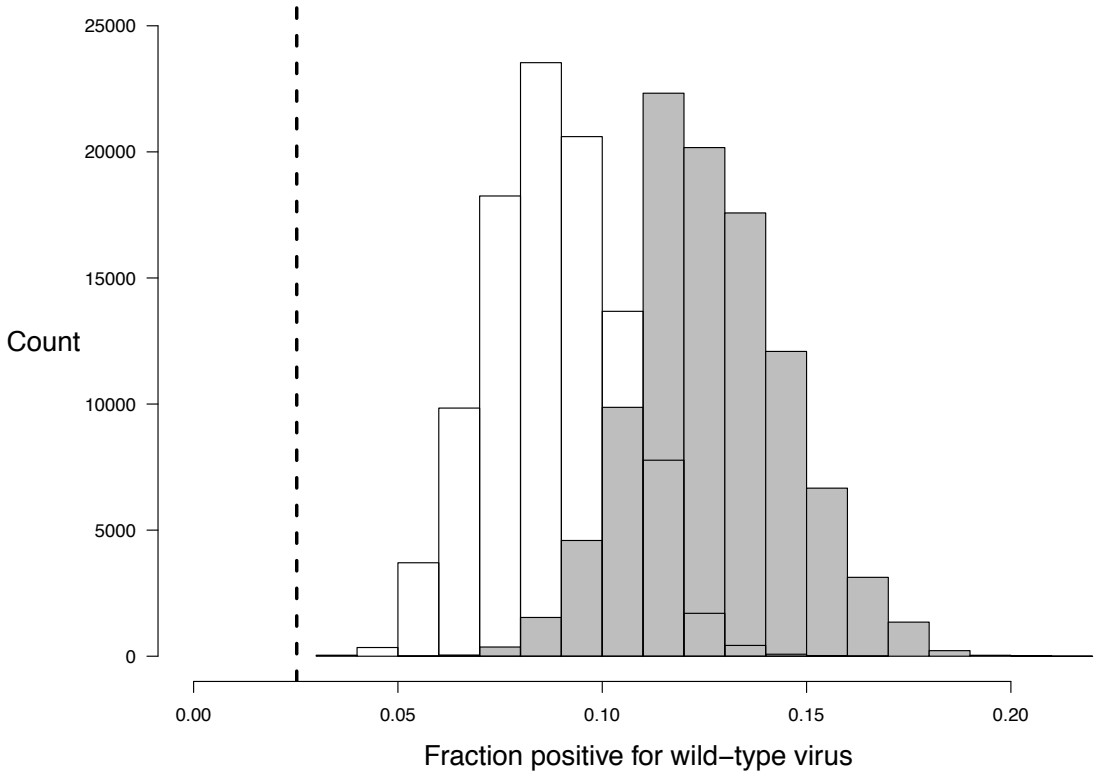

Figure S5: Bootstrap analysis results showing fraction of broiler samples that would have had "detectable virus" in a background of Rispens virus contamination similar to that seen in the layer samples. Grey bars show the distribution where "detectable virus" was defined as wild-type virus no more than 10-fold less concentrated than Rispens virus. White bars show the distribution where the threshold for detection was assumed to be equal concentrations. The dotted line is the fraction of layer samples that had detectable wild-type virus by qPCR.

A.4 Assay precision

For each of our collected dust samples, duplicate technical replicates of dust samples were weighed out and assayed independently. We used these duplicates to determine the technical error in our methods. The error in qPCR tends to occur on a log scale and so means and

variances for these data were calculated on the log scale. In practice, this was done by estimating a variance for each set of duplicates. Estimating a variance is of course impossible when virus levels are undetectable, and so we only used those samples with detectable virus. We therefore calculated sample variances for each of our sets of samples with detectable virus. We then averaged these values together and took the square root to determine our best estimate of the sample standard deviation. This analysis included 1374 sets of samples. The sample standard deviation on a log_{10} scale was determined to be 0.319 with 95% confidence intervals [0.308, 0.331]. This meant that the standard deviation on a single technical replicate was between about 2.03 and 2.14 fold.

To calculate the variation between biological replicates, we then repeated the above analysis, using our biological replicates instead of our technical replicates. That is, we calculated the sample variances between sets of samples collected from the same house of the same farm in the same collection trip. This analysis included 465 sets of samples. The sample standard deviation was determined to be 0.556 log_{10} units with 95% confidence intervals [0.513, 0.599] . This related to a standard deviation of between about 3.26 and 3.97 fold.

In a further analysis we tested whether these variances changed with the mean. This was done by regressing point-wise estimates of the standard deviation against the mean virus concentration in those samples, and performing a likelihood ratio to test whether mean virus concentration was a significant predictor of the standard deviation. In both the technical replicates ($\chi^2 = 250.0, d.f. = 1, p < 0.001$) and the biological replicates ($\chi^2 = 16.0, d.f. = 1, p < 0.001$), we observed a significant effect of the mean on the standard deviation in the data, such that the standard deviation was inversely correlated with the mean (fig. S6). For the analyses done in this paper, ignoring this heteroscedasticity would have been unlikely to have much effect. One place that this would have had a consequence was in the error bars in figs. 6 and 7, which should have been slightly larger when virus concentrations were low,

and slightly smaller when virus concentrations were high. Nevertheless, the smoothness of the data and error bars in this figure over time suggested that the assumption of normality was providing reasonable estimates.

We nevertheless further explored the relationship between the mean and variance in the technical and biological replicates by calculating means and variances for the data on the natural scale, as opposed to the log scale. Regressing the log of the variance against the log of the mean for technical replicates revealed that the slope of this relationship was 1.71 ± 0.02, suggesting that these data followed a compound Poisson-gamma distribution (fig. S7, (Jørgensen, 1987)). Note that because this was performed on a log scale, the 5 of 1374 sets of samples with an estimated variance of 0 needed to be excluded from the analysis. Repeating the analysis for our biological replicates, we found that the relationship between the log of the variance and the log of the mean was not statistically different from 2 (fig. S7, likelihood ratio test, $\chi^2 = 1.04, d.f. = 1, p = 0.31$), suggesting that these data followed a gamma distribution (Tweedie, 1985). For the purposes of the analyses in this paper, however, the assumption of normality on the log scale was probably sufficient.

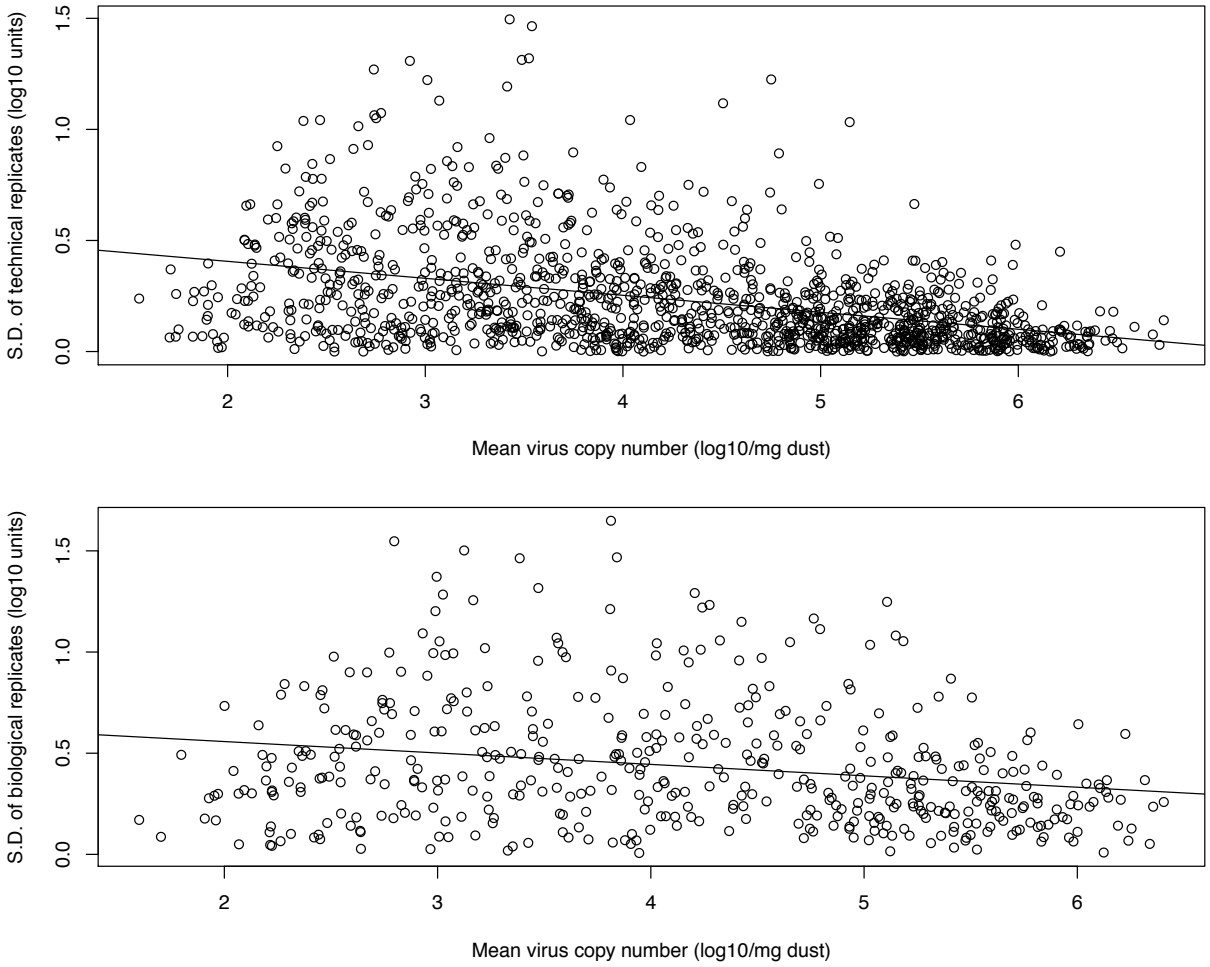

Figure S6: The relationship between the mean, and the standard deviation of technical (top) and biological (bottom) replicates using our qPCR assay. Means and standard deviations were calculated on the $\log_{10}$ scale. Solid black lines show the best fit regression line.

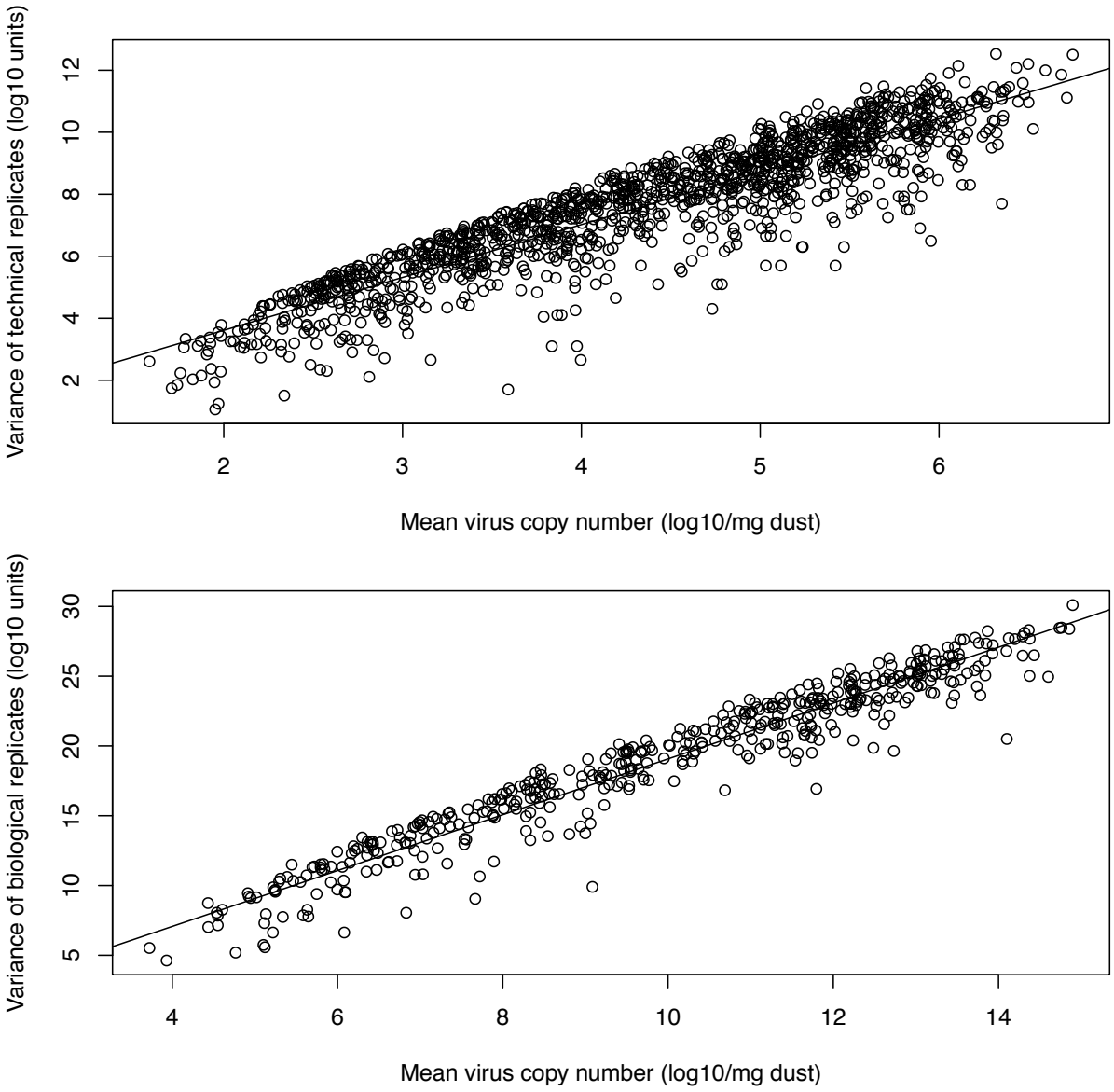

Figure S7: The relationship between the $\log_{10}$ of the mean, and the $\log_{10}$ of the variance, when mean and variance are calculated on the natural scale. The top panel shows the data from technical replicates with the solid line depicting the best fit regression line. The bottom panel shows the data from biological replicates with the solid line depicting the best fit regression line where the value of the slope is fixed at 2.

A.5 qPCR accuracy

The accuracy of our qPCR was assessed using a combination of dilutions and mixtures. First, we wanted to determine whether our qPCR was able to accurately quantify virus copy number over a wide range of virus densities. We did this by creating a set of samples from serial ten-fold dilutions of our DNA standards. Our highest concentration was 2.41×10^7 virus genome copies per 4 μl, and our lowest dilution was 2.41×10^0 virus genome copies per 4 μl. With the exception of the highest density, in which target DNA is substantially higher than any of our field samples, and lowest density, which was below our limit of detection, these dilutions followed an observed linear decline in Ct of 3.39 per 10-fold dilution. The decline in Ct per 10-fold dilution can be used to calculate the amplification efficiency of a qPCR assay with a 100% efficiency relating to a decline of 3.32. Our observed curve related to an amplification efficiency of 97.1% with an R^2 of 0.999. In practice, each plate of qPCR reactions contained a similar dilution series used as a standard for quantification. From these dilution series, we found a median linear decline in Ct of 3.34 per 10-fold dilution, with 95% of assays between 3.23 and 3.44. This related to an amplification efficiency of 99.4%, ranging between 95.3% and 104.0%. In these assays, our median limit of detection was 55.0 virus copies per mg of dust, with 95% of assays between 21.4 and 168.6.

We next wanted to test whether nontarget contaminating DNA would interfere with the accuracy of the assay. In unpublished data, we saw that in DNA extracted from dust, nontarget DNA was more prevalent than wild-type virus DNA at a ratio of approximately 5000 to 1 by mass. We therefore mixed plasmid DNA of the virus used as our standard with plasmid DNA of the chicken ovotransferrin (*ovo*) gene to create samples with 500:1, 5000:1, and 50000:1 ratios of chicken DNA to wild-type DNA by mass. In practice, we mixed a concentration of 1000 copies of the virus plasmid per 4μl with the necessary quantity of the plasmid having chicken *ovo* gene insert, and this was performed three times for each target

concentration. These samples then underwent serial 10-fold dilutions to generate samples with 100 and 10 copies of the virus plasmid per 4μl. Our typical standards were run with these samples to ensure that the assay remained accurate across these conditions. We found that contamination from nontarget DNA added very little bias to our estimates (fig. S8).

We next attempted to validate our quantification of virus from dust by mixing virus positive dust with virus negative dust and comparing the resulting virus concentrations measured by qPCR to the values expected. We used 8 different samples of virus positive dust that spanned a range of virus concentrations, and we mixed this with a dust sample that came from a farm in which virus was never detected in any of our samples. Positive dust and negative dust were mixed to generate ratios of positive to negative 1:0, 10:1, 3:1, 1:1, 1:3, 1:10, 1:30, and 1:100. These mixtures were then processed in an identical manner to our field collected samples. We analyzed the results using linear models. These models took the form:

$$\mathrm{lm}(V \sim 0 + F + F:S), \tag{1}$$

$$\mathrm{lm}(V \sim 0 + F:S). \tag{2}$$

Here we use the "R" syntax for linear models. V is the measured concentration of virus per mg of mixed dust, F is the fraction of dust that came from the virus positive sample, and S is a factor denoting the sample of virus positive dust used. "$F:S$" is used to express an interaction term between F and S, and "0" is used to express that the intercept of the model is forced through 0. In eq. (1), "F" is treated as a factor when it is on its own. If a likelihood ratio test were unable to reject the model described by eq. (2), that would suggest that the data followed the expected dilution pattern. This was exactly what we found (fig. S9, likelihood ratio test, $\chi^2 = 10.5, d.f. = 7, p = 0.161$), confirming the validity of our DNA extraction protocol.

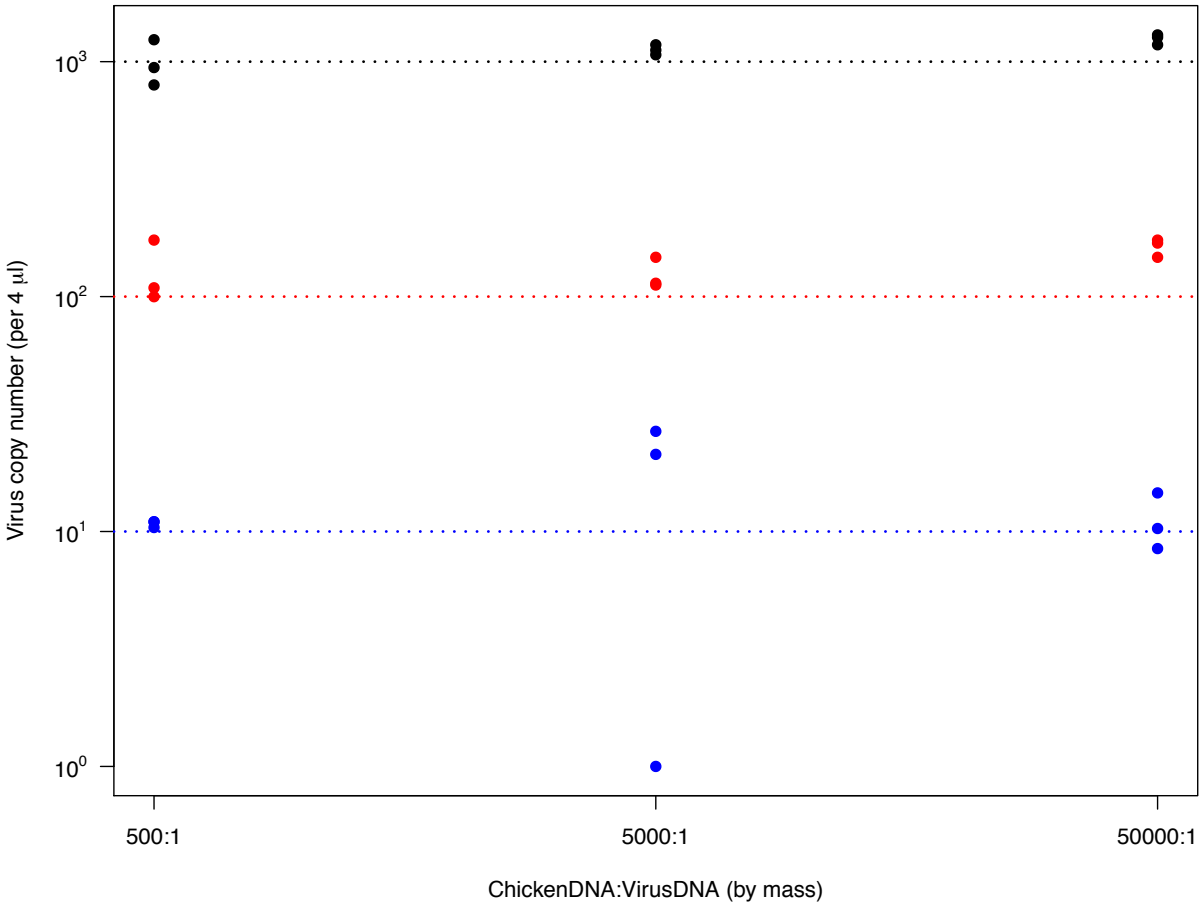

Figure S8: The effect of nontarget DNA (chicken plasmid DNA) presence on virus quantification. Each point shows the virus copy number measured by qPCR. Expected virus quantities for each point are shown by the color-matched dotted lines. These data show that over a wide range of non-target DNA ratios, quantification of virus DNA was unaffected by contaminating DNA.

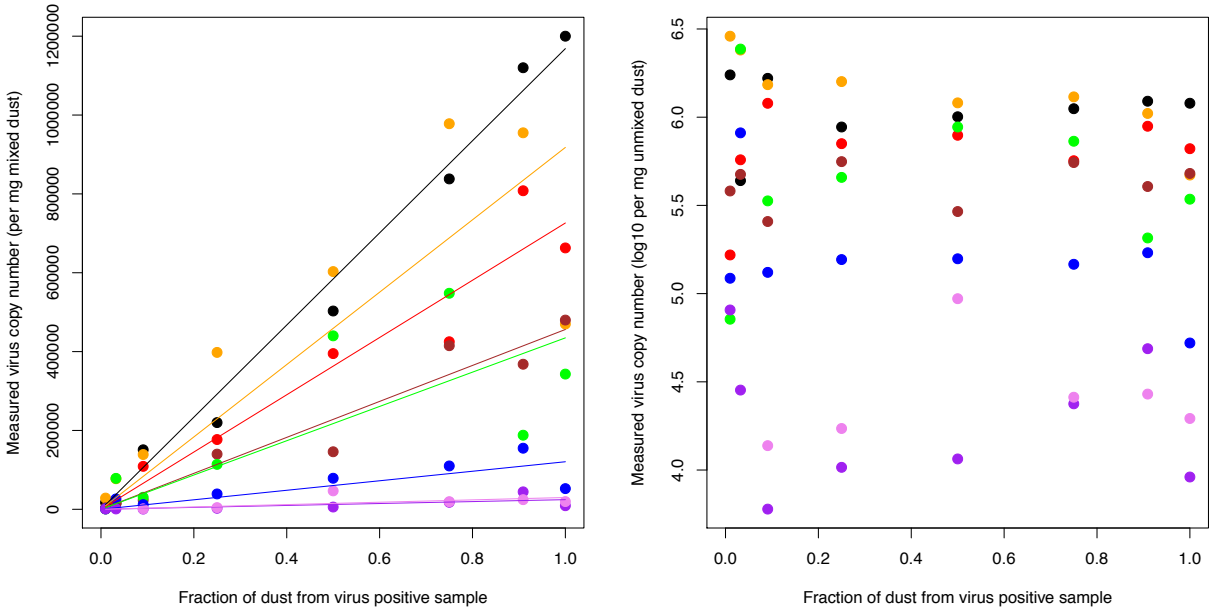

Figure S9: The effects of dust mixing on virus quantification. Panel A shows the measured virus copy number that resulted from the pre-DNA-extraction mixing of virus positive and virus negative dust. Colored points show data that resulted from using different samples of virus positive dust. Color-matched lines are the regression lines that resulted from fitting eq. (2). Panel B shows back calculated estimates of virus copy number per mg of dust in the unmixed virus positive dust. This value was calculated by dividing virus copy number in the mixed dust by the fraction of dust that was from the virus positive sample. The general flatness of color-matched points across the range of mixtures validated our claim that virus copy number can be assessed from dust samples.

A.6 Feather tip data analysis

A complication in analyzing the feather tip data was accounting for background virus contamination. As shown from the air tube data, in some chicken houses, virus DNA was sufficiently prevalent that it was detectable even without being intentional introduced. To account for this potential contamination, control tubes were used during the collection of feather tips. Each set of feather collections were associated with a "short open" tube, which was left open for the amount of time that it took to collect a single feather from a single

bird (generally less than 10 seconds), and a "long open" tube, which was left open for the amount of time that it took to collect two feathers from each of ten birds (approximately 30 minutes). This latter tube was open substantially longer than the former because of the time associated with bird selection and capture. This control was therefore likely to be a substantial overestimate of potential virus contamination.

To determine the fraction of birds with detectable virus in feather tips, we compared the qPCR virus copy number measurement of sample tubes to that of the control tubes. If at least one of the two feather samples from a bird was higher than both control tubes, that bird was considered to have detectable virus in its feather tips, and otherwise not. Of course, had we chosen a different cutoff our estimates would have differed. Fig. S10 shows this, by using a more conservative requirement that both feather samples needed to give higher copy number reads than both controls, and a less conservative cutoff that at least one of the two feather samples needed to give a higher copy number read than the "short open" control.

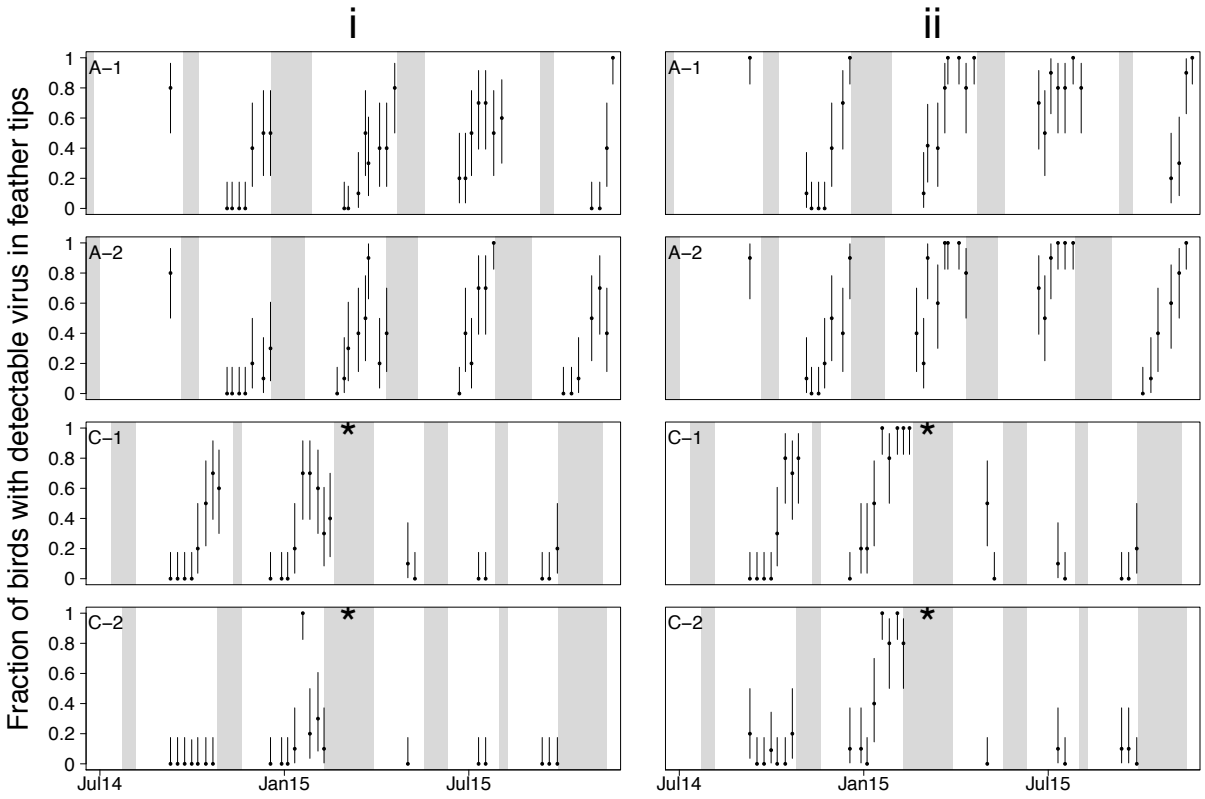

Figure S10: Fraction of birds with detectable virus in feather tips. Column i shows the data when using the highly conservative requirement that both feather samples must yield higher virus copy number estimates than both controls to be considered a virus positive bird. Column ii shows the data using a less conservative requirement that at least one feather must yield a higher virus copy number than the "short open" control. Together these two sets of plots bound our uncertainty due to potential contamination.

A.7 Virus detectability

We used virus shed rates from the published literature and data from our surveillance to estimate how many chickens would need to be infected before the virus concentration in the house would be sufficiently high for our assay to detect its presence. The concentration of virus infected dust in a house at the end of a cohort can be described by the following

equation.

$$C = \frac{(T-\tau)}{T}\frac{(1-\phi)VI}{N}. \tag{3}$$

Above, C is the concentration of virus per mg of dust, T is the duration of a cohort, τ is the incubation period between infection and maximum virus shedding, ϕ is the fraction of house dust that is produced by non-chicken sources such as litter or feed, V is the amount of virus produced per mg of dust from an infected chicken, I is the number of birds infected, and N is the total number of birds in the chicken house. This is of course only a rough approximation. The true equation would depend on a multitude of unknown variables including the breed specific virus shed rate of the chickens, and the virulence rank of the virus. Nevertheless, this calculation is a useful starting point for interpreting our limit of detection.

In the above equation, we made several assumptions. First, we assumed that virus was well mixed. Our analysis has already shown that the virus was not perfectly mixed within dust samples, but the samples were only moderately variable (see Appendix A.4) suggesting that it would have little impact on our detection limit. Second, we assumed that virus infection happened immediately after placement. This assumption seems reasonable given that the virus is highly infectious and resistance increases with bird age (Sharma et al., 1973; Witter et al., 1973). Third, we assumed that dust samples were collected at the end of a cohort. In fig. S11, we showed how cohort duration affected our ability to detect virus. This figure could also have been interpreted as our sensitivity to detect virus if collected at the time on the x-axis, although note that this would underestimate our ability to detect virus, making it overly conservative. Fourth, we assumed that the cohort to cohort infection level was at an equilibrium value. This assumption was necessary to account for dust that persisted between flocks. In our experience, however, for most farms, very little dust was held over between flocks, and so this assumption was probably unnecessary. Fifth, we assumed

that ventilation rates were negligible. If this assumption were violated, it would make our analysis more conservative, because dust produced early would be more likely to be removed by ventilation than dust produced late, and early dust is likely to have less virus than late dust.

Given these caveats, we solved the above equation for I to determine how many infected chickens would be needed to detect virus in a chicken house.

$$I = \frac{CNT}{V(1-\phi)(T-\tau)}. \tag{4}$$

We can then substituted in values for C, V, N, ϕ, T, and τ to calculate I. In practice we used a conservative estimate of the limit of detection $C = 10^2$ (Appendix A.5), two different estimates of V from bivalent vaccinated birds (Islam and Walkden-Brown, 2007), a higher end estimate of the number of chickens in the broiler houses we visited $N = 30,000$, and the incubation period before maximum shedding $\tau = 21$ from Islam and Walkden-Brown (2007). For the fraction of dust that came from non-chicken sources ϕ, we multiplied the fraction of dust not attributable to feed for caged layers, 0.15, by the fraction of dust not attributable to litter for layer birds raised on bedding material, 0.385 (Collins and Algers, 1986), and subtracted this product from the number one. This gave us the estimate of $\phi = 0.94$. We used a range of values for T to show how the duration of a cohort would affect these results. Using these numbers, our limit of detection was estimated to be between 18.1 and 53.6 infected birds per cohort in a broiler house.

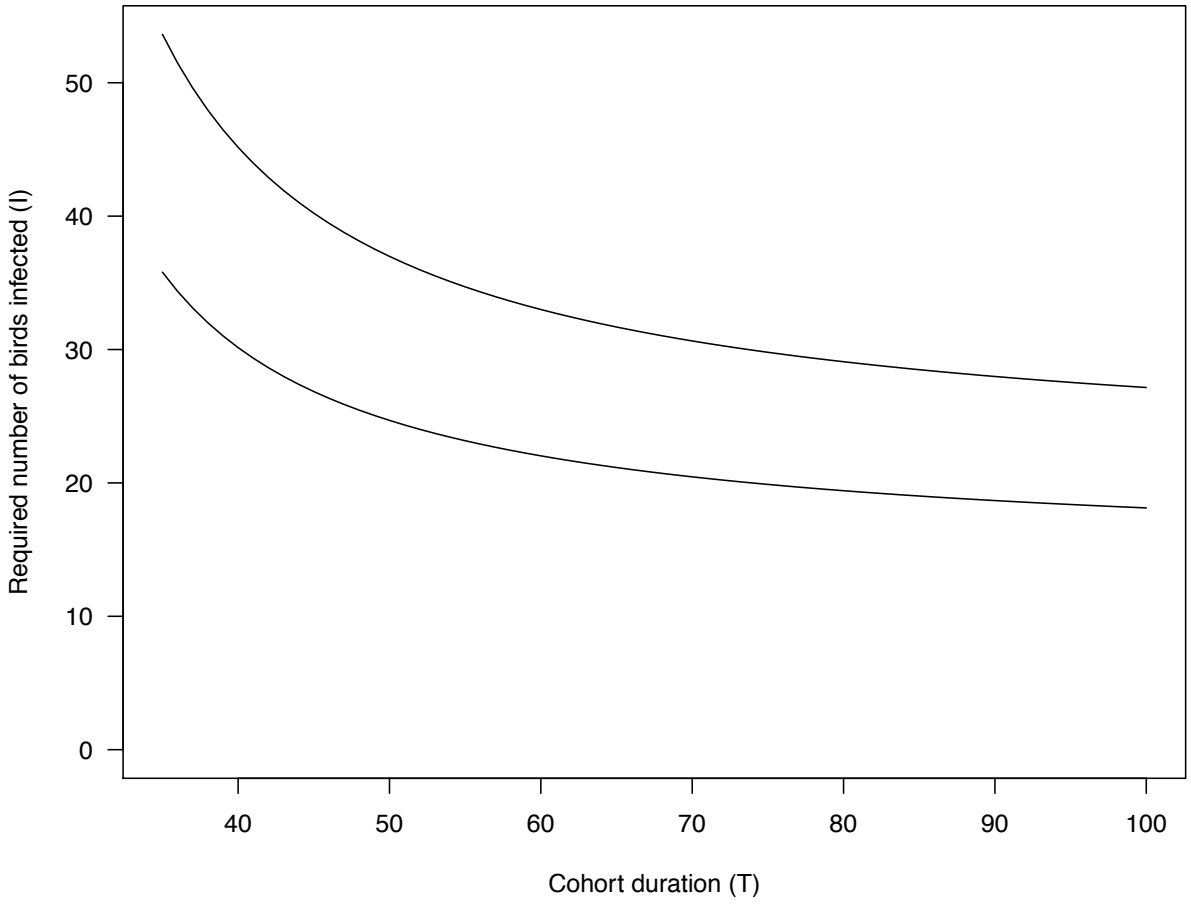

Figure S11: Estimated number of birds that would need to be infected to detect virus by our qPCR assay. The two different lines show detection limits for the two different virus strains characterized by Islam and Walkden-Brown (2007). For both cases, over a wide range of cohort durations, our assay was sufficiently sensitive to detect virus when even a small number of birds were infected.

A.8 Sensitivity to priors

As mentioned, our MCMC chains were only able to converge when we supplied informative priors. We explored whether our results were driven by our choice of prior by rerunning our

analysis with three different sets of priors. In the main text we used the parameters scale $V = 3$ and degrees of freedom $\nu = 5$ for random effects, and the parameters mean $\mu = 0$ and variance $\sigma^2 = 49$ for fixed effects (prior set "A"). We explored the effect of changing the random effects prior to $V = 5$, $\nu = 1$, while leaving the fixed effects prior unchanged (prior set "B"), and we explored the effects of changing the random effects prior to $\mu = 0$, $\sigma^2 = 16$ (prior set "C"). All three sets of priors are shown in fig. S12. The results of our analyses were similar for all three sets of priors, and the model ranks according to DIC were unchanged. This provided evidence that our conclusions are robust to our choice of prior.

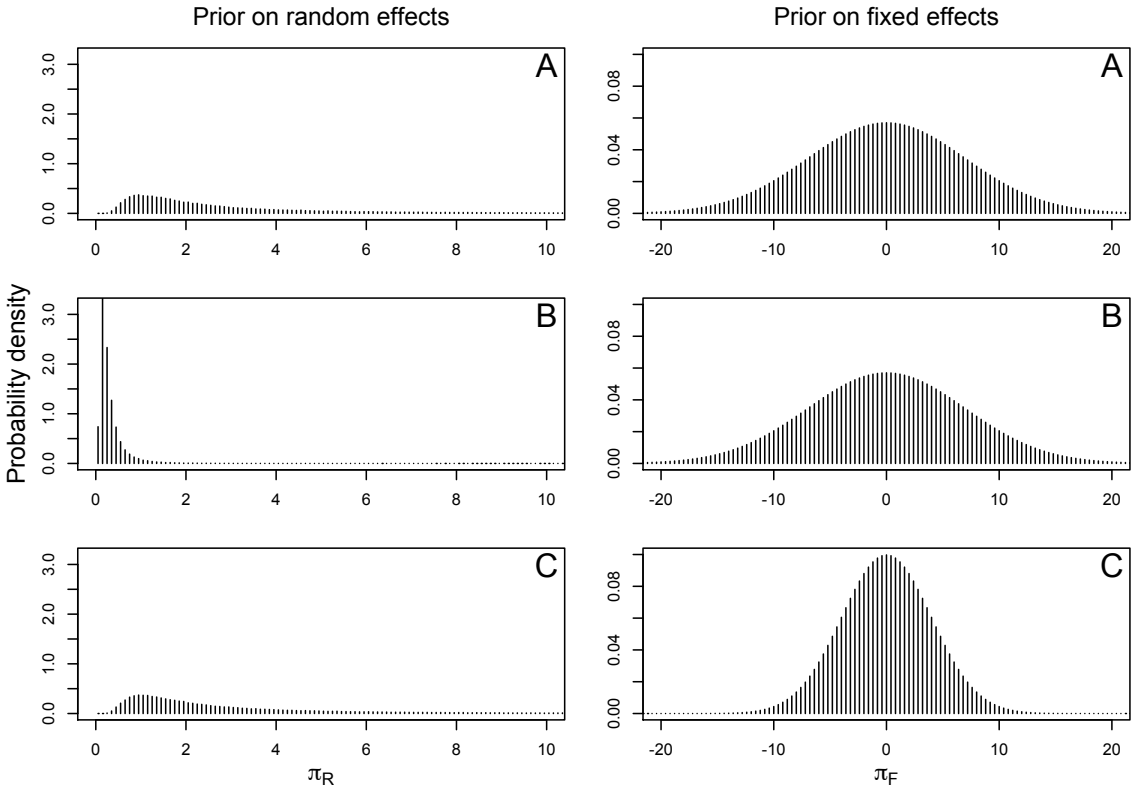

Figure S12: Three sets of priors explored, labelled "A", "B", and "C", with the priors for random effects on the left, and the priors for fixed effects on the right. "A" is the set of priors used in the main text.

A.9 Summary plots of marginal incidence

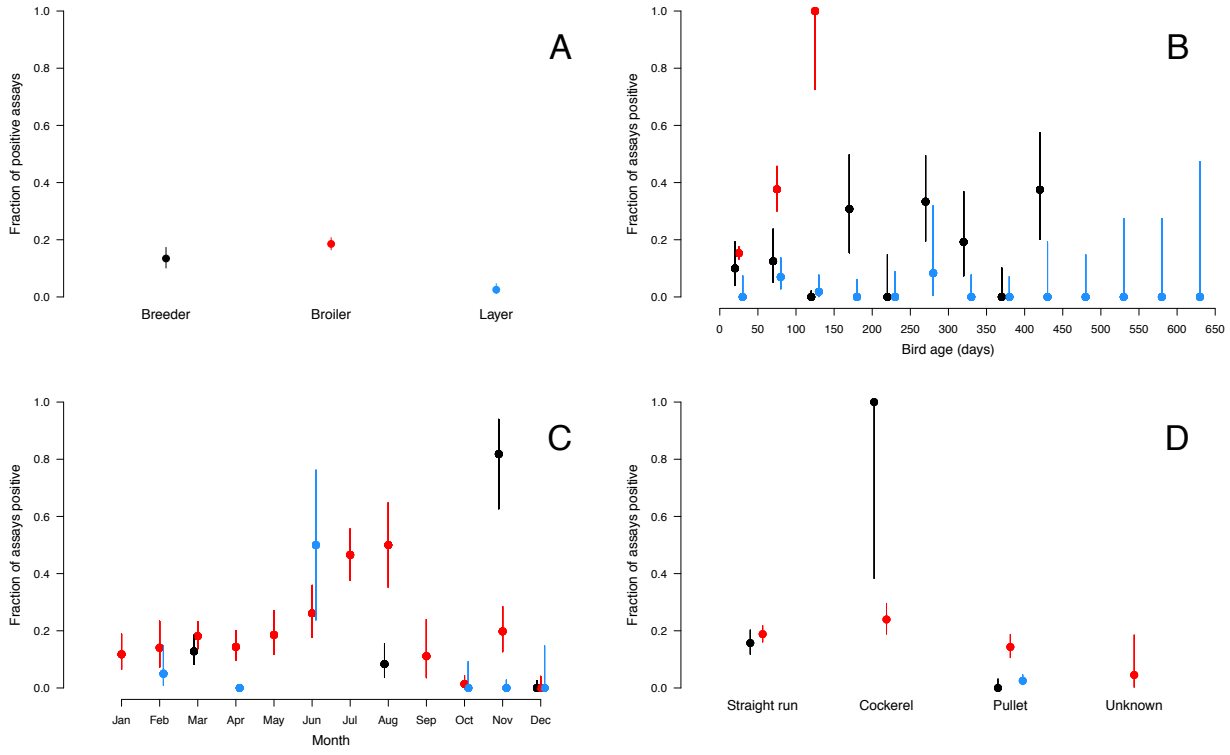

Figure S13: Summary plots depicting the fraction of assays that were positive as a function of production type (A), bird age (B), month of the year (C), and sex (D). As in fig. 2, black denotes breeder facilities, red denotes broiler facilities, and blue denotes layer facilities. Error bars are 95% confidence intervals on the estimate of the fraction of positive assays. Note, however, that these error bars were generated under the assumption that the data were perfectly binomially distributed data, which was almost certainly untrue for these data because they were marginalized over other factors that we showed to be important to virus prevalence. For these reasons, we were cautious not to over interpret these plots.

A.10 Estimating virus concentrations and 95% confidence intervals

Confidence bounds for all prevalence data were calculated using likelihood. 95% intervals spanned the range estimates that were within 1.92 log likelihood units of the maximum

likelihood estimate.

For our quantitative data, the standard practice of using sample standard errors to generate confidence intervals on data points was impossible, because samples were routinely below the qPCR limit of detection. We were therefore forced to estimate virus concentrations and 95% confidence intervals in a slightly more complex way. First, we calculated the standard deviation of our samples as described in subsection Assay repeatability. We then used this point estimate of the standard deviation in the following way.

There were two possible outcomes when running our qPCRs, we either observed a virus concentration ($X = x$), or the virus concentration was so low that it was undetectable ($X < C$) where C was our limit of detection. For our calculations below, we assumed a limit of detection of $C = 100$ target copy number per mg of dust which was a slightly conservative estimate of our limit of detection in qPCR runs (Appendix A.5). In cases where virus was quantifiable by qPCR, we assumed the likelihood of the data on a log scale followed a normal distribution. These values were calculated in the R programming language using the function 'dnorm'. In cases where virus could not be quantified by qPCR, we extended this methodology by using a Bernoulli distribution for the likelihood function, where the probability of not detecting virus was exactly equal to the cumulative distribution of a normal distribution up to the limit of detection $log_{10}(C)$. In R, these values were calculated using the function 'pnorm'. The sum of the log of these values was a likelihood function. We thus generated a maximum likelihood estimate of virus concentration by maximizing this likelihood function using the R function 'optimize'. Upper and lower confidence intervals for virus concentrations were then calculated by solving for the values of virus concentration for which the likelihood function was 1.92 log likelihood points worse than the maximum likelihood. This cutoff was chosen so that we would have 95% confidence intervals. Note that the data did not exactly follow the normal distribution (Appendix A.4). In practice, this meant that our estimates of the confidence intervals were slightly inflated for points at

high virus concentrations, and slightly deflated for points at low virus concentrations.

A.11 Longitudinal cohort spline

To look for effects of cohort age on virus concentration, we fit cubic smoothing splines to the longitudinal data on virus concentration. We treated data from each farm individually, but within each farm, all data across houses, flocks, and samples were used to infer the effect of cohort age. As mentioned in the methods, we used 4 knots, but we explored every number of knots up to 9 and saw nearly identical results. For this analysis, we fit the splines on a log scale, and we treated samples that were negative by qPCR first as values of 0, and second as values at our qPCR limit of detection. These two sets of analyses gave qualitatively similar outcomes. In fig. S14 we show the results for 4 knots, treating negative samples as though they were at a value of 10^2 virus copies per mg of dust. Note the decrease in virus concentration that occurred early in cohorts, and the subsequent increase that occurred as cohorts aged.

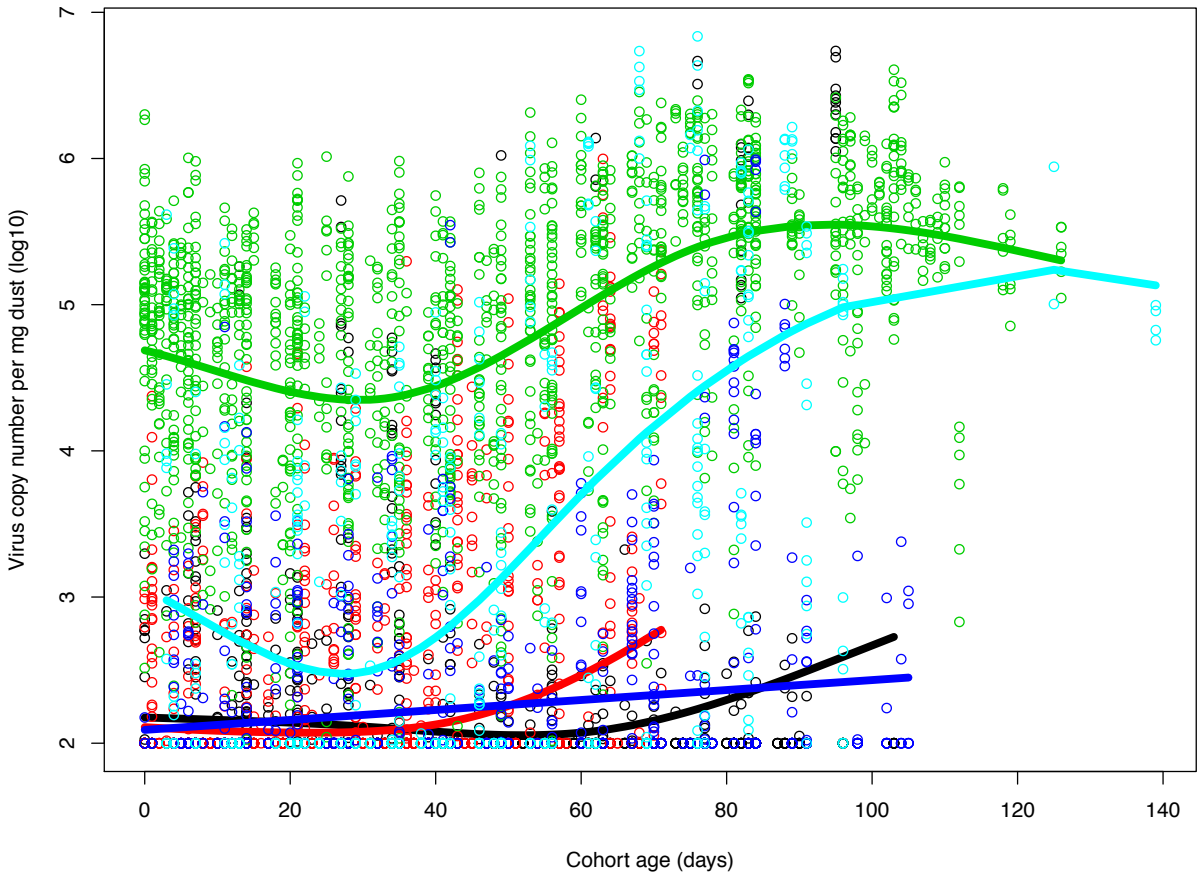

Figure S14: The effect of cohort age on virus concentration per mg of dust. All data points from our longitudinal surveillance are plotted as open circles with different colors representing data from different farms. The smoothing splines are plotted as solid lines matching the color of the points for the respective farms.

A.12 Autocorrelation within houses

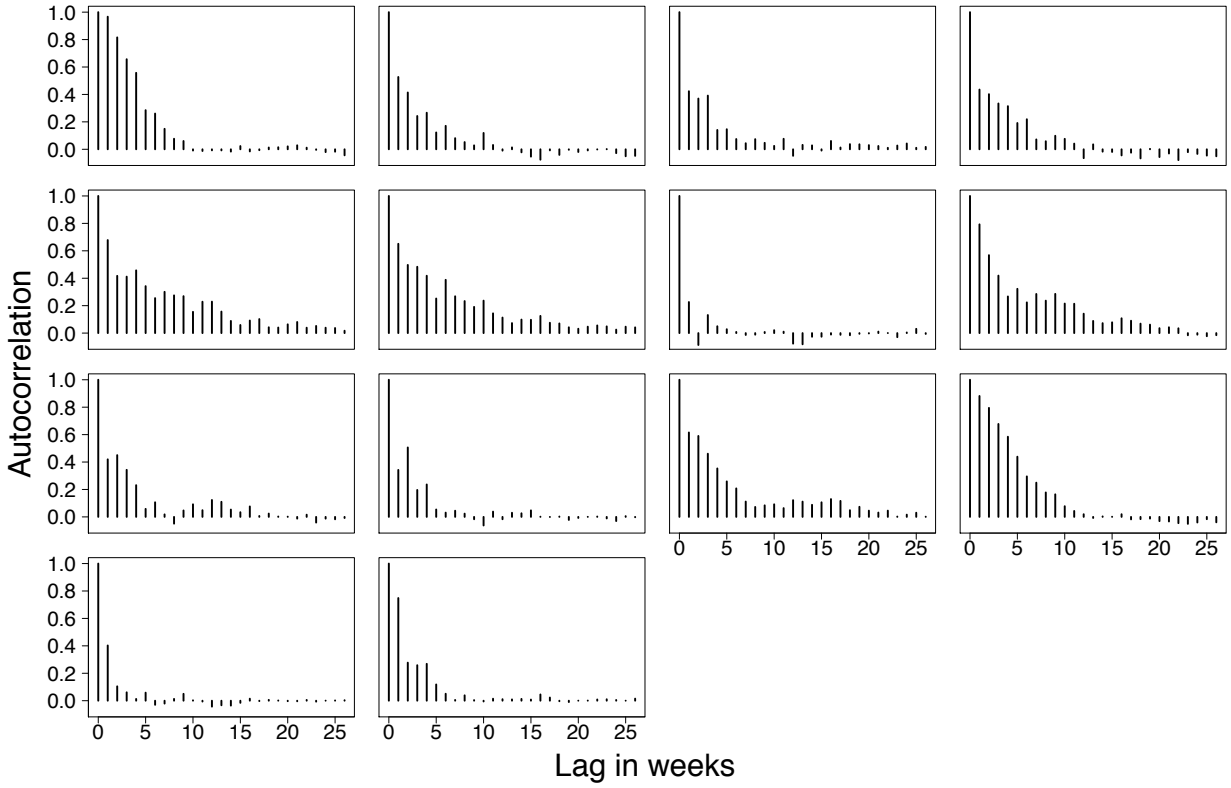

Figure S15: Plots showing the autocorrelation in virus concentration in the longitudinal data. Each plot shows data from a different house.

A.13 Sample humidity or qPCR inhibitors

In the main text, we stated that the "U" shape pattern observed in virus concentration within a cohort iwass likely to be driven either by dust dilution or virus DNA degradation followed by re-concentration. As an alternative to these biological explanations, we explored whether this pattern could be caused by either of two technical factors. First, water content in dust samples might have changed over the course of a cohort resulting in substantial changes to the dry mass of the dust included in DNA extractions. Second, qPCR inhibitor concentration in dust might have similarly changed over the cohort limiting our ability to

detect virus DNA when present. Here, we show that neither of these explanations can explain the observed pattern in the data.

To explore whether changing water content could explain the "U" shape pattern in the data, we used dust samples collected from the fourth and fifth cohorts of farm "A", house 2, and from the eighth cohort of farm "A", house 2 (fig. 6). These cohorts were chosen because their dynamics were among the most and least extreme fluctuations in virus concentration. For each collection trip in each of these flocks, we randomly chose one representative dust sample. A 2 mg aliquot of this dust sample was used in our standard qPCR assay to re-determine virus concentration. In addition, a 10 mg aliquot was transferred into a clean 1.5 ml Eppendorf tubes. As controls, we had one tube with 10 mg of previously desiccated dust, one tube with 21 μl of distilled water, and one tube with both 2.5 mg of desiccated dust and 9.8 μl of water. All tubes were opened and placed on a VWR Analog Heatblock for 5 hours at 37 °C. This was sufficiently long for all visible water to evaporate from control tubes. The mass of the controls changed as expected, suggesting that our protocol was sufficient to desiccate the samples. The 10mg dust samples on average lost 1.2 mg of mass through desiccation, with a range of -0.2 mg to 6.8 mg. We divided the mass remaining after desiccation by the original mass to determine the fraction of each dust sample that was dry mass. We then divided the virus concentration determined through qPCR by these fractions to explore whether this would eliminate the "U" shape in our data. We found that the dynamics were nearly unchanged (fig. S16), meaning that water content alone could not explain the pattern observed in the data.

To explore whether qPCR or DNA extraction inhibitors might have created this "U" shape pattern in the data, we mixed together pairs of samples with high virus concentration and low virus concentration. Because we were interested in whether low virus concentration samples contained inhibitors, we mixed these samples at a ratio of 9:1, such that an 18 mg sample with low virus concentration was combined with a 2 mg sample with high virus

concentration. If true virus dynamics in a house were essentially flat, and the pattern observed was due to inhibitors, then mixing these samples together should result in virus concentrations consistently lower than expected. Alternatively, if the virus concentration truly changed over this time period, these mixtures would have given qPCR reads similar to those expected by simple averaging. We found the latter to be true (fig. S17), suggesting that inhibitors were unlikely to drive the virus dynamics seen in the data. We were therefore left to conclude that the dynamics seen in the data were probably caused by biological, rather than technical, factors.

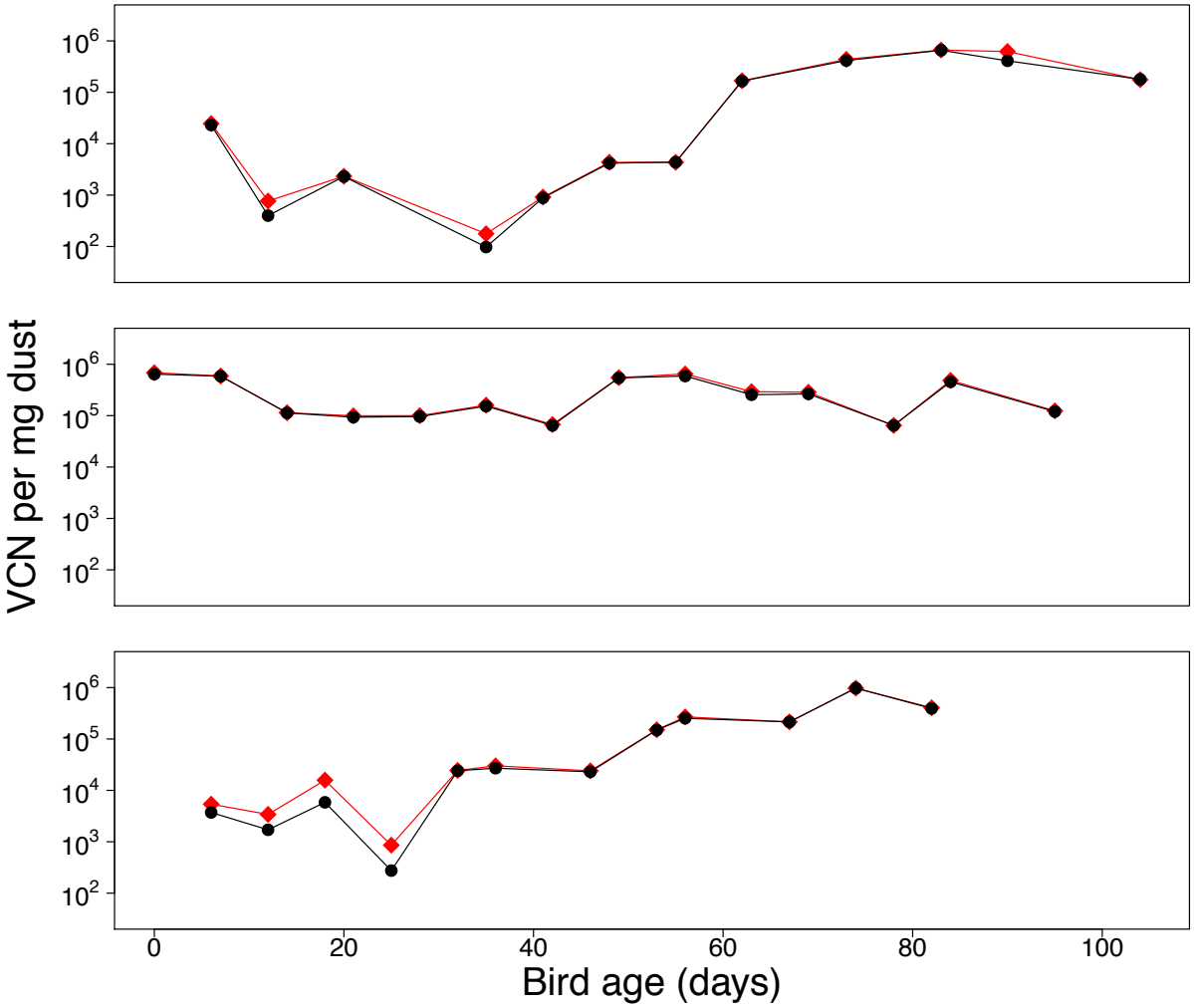

Figure S16: Virus copy number per mg of dust and per mg of desiccated dust. The three plots show data from three different flocks as a function of bird age. Black circles show virus copy number per mg of dust. Red diamonds show the data corrected to be virus copy number per mg of dry mass. The values differed only slightly, and thus we thus concluded that water content was unable to explain the "U" shape seen in virus concentration within cohorts.

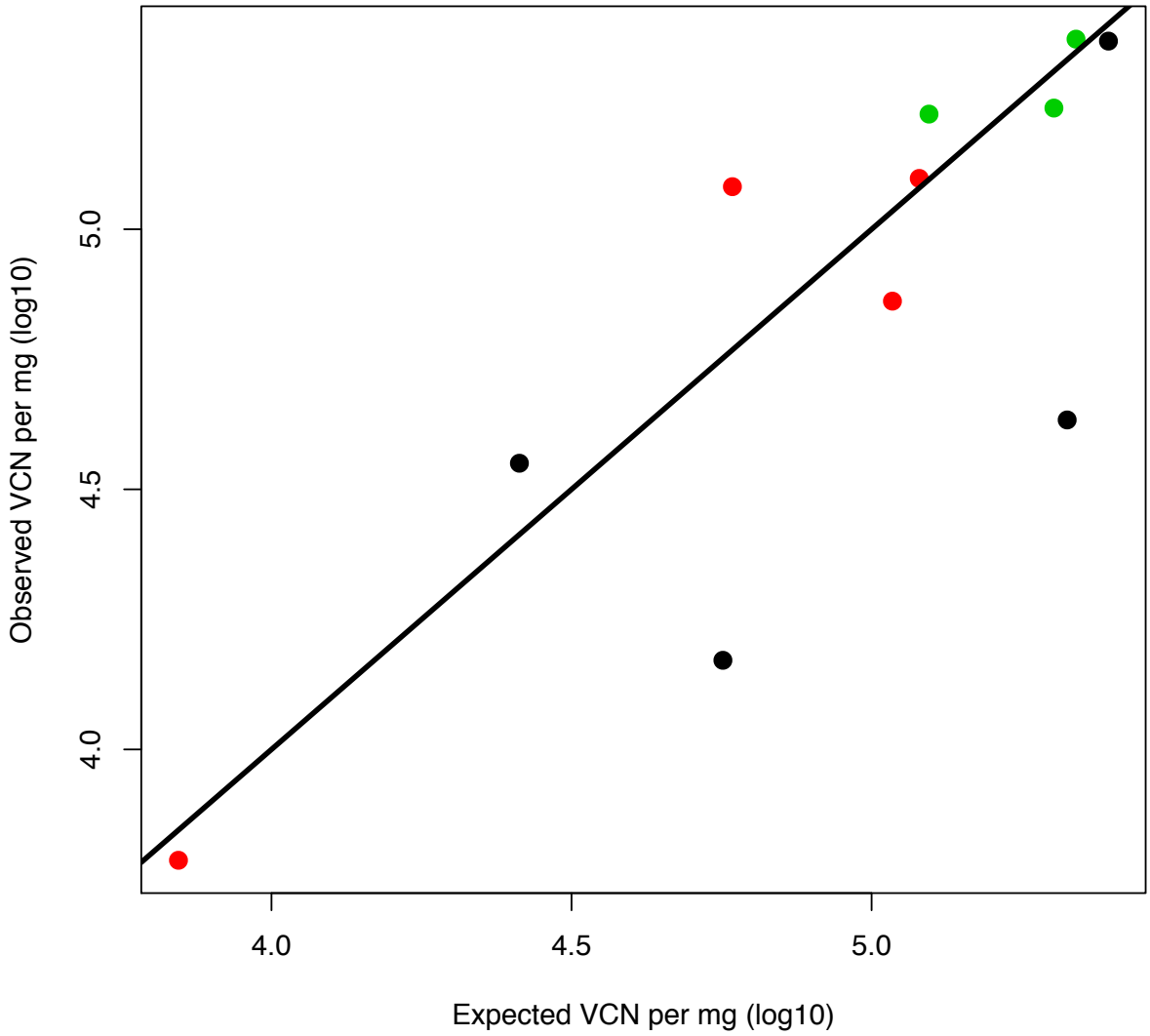

Figure S17: Expected vs. observed virus concentration in mixed samples. Each point shows a mixture of a different pair of samples. Each color depicts samples collected from different flocks. The solid black line is a one to one line.