1. Supplementary materials

1.1 *Hellinger Distance*

In order to capture significant evolutionary changes within the genomes of the viral populations, we need a measure for quantifying the changes. We chose an f-divergence, the Hellinger distance, H, to measure the similarity between two probability distributions. Under Lebesgue measure, for two probability density functions f and g, the squared Hellinger distance can be expressed as following:

$$H^2(f,g) = \frac{1}{2}\int \left(\sqrt{f(x)} - \sqrt{g(x)}\right)^2 dx = 1 - \int \sqrt{f(x)g(x)}dx. \tag{1}$$

The Hellinger distance is a metric. The larger H is, the more different f and g are.

We prefer the Hellinger distance over relative entropy, the Kullback-Leibler divergence (KL), because symmetry is a desired property for the comparison of distributions. One can also use a symmetrised KL, such as the Jensen-Shannon divergence. We used the Hellinger distance to compare two marginal posterior distributions of the probability parameters given all cluster assignments and every read count, $P_{k_1}|c_1,\cdots,c_N,Y_1,\cdots,Y_N$, and $P_{k_2}|c_1,\cdots,c_N,Y_1,\cdots,Y_N$. The distance measures how similar the two allelic positions or same allelic position at two different time points are. Applying the squared measure (1) to two marginal posteriors, we have

$$H^2(P_{k_1},P_{k_2}|c_1,\cdots,c_N,Y_1,\cdots,Y_N) = 1 - \frac{B(\vec{\beta}_{k_1,k_2})}{\sqrt{B(\vec{\alpha}_{k_1})B(\vec{\alpha}_{k_2})}}, \tag{2}$$

where $\vec{\alpha}_i = \left(\alpha_i^1,\cdots,\alpha_i^J\right)$; $\vec{\beta}_{k_1,k_2} = \left(\frac{\alpha_{k_1}^1+\alpha_{k_2}^1}{2},\cdots,\frac{\alpha_{k_1}^J+\alpha_{k_2}^J}{2}\right)$; $B(a^1,\cdots,a^J) = \frac{\prod_{j=1}^J \Gamma(a^j)}{\Gamma\left(\sum_{j=1}^J a^j\right)}$.

To better visualize Hellinger distances for the viral data we further applied a monotonic transformation on H: $f(H) = \ln\left(1-\ln\left(1-H^2\right)\right)$. With the definition (2) the Hellinger distance can then be transformed into

$$Ht(P_{k_1},P_{k_2}|c_1,\cdots,c_N,Y_1,\cdots,Y_N) = \ln\left(1-\ln\left(\frac{B(\vec{\beta}_{k_1,k_2})}{\sqrt{B(\vec{\alpha}_{k_1})B(\vec{\alpha}_{k_2})}}\right)\right). \tag{3}$$

1.2 *Preprocess procedure*

Continuing with the toy example in Figure **??**, the first step is to combine the datasets collected at different time points and consolidate the invariant read sites (Table 1). Assume that in the toy example $J = 5$. The five possible reads are A, C, G, T, M, as in Table **??**.

[Table 1 about here.]

The three small tables in the first row of Table 1 show the read counts obtained at time points $t_1,\ t_2$, and t_3; the second row table shows the combined data of the first row produced by merging all the sites with a particular homogeneous read type. The first few columns of the joint data (second row table in Table 1) are the consolidation of columns with single read type A, C, G, T, M, respectively. The sites with multiple read types (non-invariant sites) are copied to joint data matrix after all the combined invariant sites ($Y_1^{all\ t}, Y_2^{all\ t}, Y_3^{all\ t}$ in the toy example). In particular, $Y_1^{all\ t}$ in the joint data matrix (second row in Table 1) is formed by merging columns $Y_1^{t_1}$ and $Y_1^{t_2}$. Similarly, $Y_2^{all\ t}, Y_3^{all\ t}$ are formed from the sites that have a homogenous read of C and T, respectively:

$$Y_2^{all\ t} = Y_3^{t_1} + Y_3^{t_2}, \quad Y_3^{all\ t} = Y_2^{t_1} + Y_5^{t_1} + Y_2^{t_2} + Y_5^{t_2} + Y_2^{t_3} + Y_4^{t_3} + Y_5^{t_3}.$$

The following columns in the second row are

$$Y_4^{all\ t} = Y_4^{t_1},\ Y_5^{all\ t} = Y_4^{t_2},\ Y_6^{all\ t} = Y_1^{t_3},\ \ldots$$

The exact mapping is shown in the third row of the table. Note that this preporcessing step consolidates invariant sites and reduces the dimensionality of sequencing data without losing any significant information.

1.3 *The statement and proof of Theorem 1*

THEOREM 1: *Suppose that* $Y = [Y_1,\ Y_2]$, Y_1 *and* Y_2 *are* $J \times 1$ *random read count vectors.* $J \in \{2, 3, 4, \cdots\}$. *Further assume that* $Y_i | c_i, \boldsymbol{P} \overset{indep}{\sim} Multinomial\left(m_i; P_{c_i}\right)$, *for* $i = 1, 2$.

If $c_1 = c_2 = 1$ *and* m_i*'s are sufficiently large, then the marginal posterior likelihood ratio of*

assigning different labels over current state goes to zero almost surely,

$$\textit{i.e. } LR = \frac{\pi(c_1 = 1, c_2 = 2|Y)}{\pi(c_1 = 1, c_2 = 1|Y)} \to 0, \ \textit{a.s.}$$

Suppose there are only two genome positions to be clustered, $Y = [Y_1,\ Y_2]$. If they share the same probability parameter, then the likelihood of the two share the same parameter is one when the numbers of observations at the two sites m_1 and m_2 are large. Fix a $J \in \{2, 3, 4, \cdots\}$. Without loss of generality, assume $c_1 = c_2 = 1$, and then the marginal posterior likelihood ratio of splitting the two over current state on the log scale is the following:

$$\begin{aligned} LR &= \log(\pi(c_1 = 1, c_2 = 2|Y)) - \log(\pi(c_1 = 1, c_2 = 1|Y)) \\ &= \sum_{j=1}^{J} \left[\log\Gamma(y_1^j + J^{-2}) + \log\Gamma(y_2^j + J^{-2})\right] - \log\Gamma(m_1 + J^{-1}) - \log\Gamma(m_2 + J^{-1}) \\ &\quad - \sum_{j=1}^{J} \log\Gamma(y_1^j + y_2^j + J^{-2}) + \log\Gamma(m_1 + m_2 + J^{-1}) \end{aligned}$$

Claim: $LR \longrightarrow -\infty \ \ a.s.$

Proof. Recall that Stirling's formula provides the following approximation:

$$\log\Gamma(z) \approx \frac{1}{2}\log(2\pi) - \frac{1}{2}\log z + z\log z - z$$

Therefore,

$$\begin{aligned} LR &\approx \sum_{j=1}^{J} \Big[\frac{1}{2}\log(2\pi) - \frac{1}{2}(y_1^j + J^{-2}) - \frac{1}{2}(y_2^j + J^{-2}) + \frac{1}{2}(y_1^j + y_2^j + J^{-2}) + (y_1^j + J^{-2})\log(y_1^j + J^{-2}) \\ &\quad + (y_2^j + J^{-2})\log(y_2^j + J^{-2}) - (y_1^j + y_2^j + J^{-2})\log(y_1^j + y_2^j + J^{-2}) - J^{-2}\Big] - \frac{1}{2}\log(2\pi) \\ &\quad + \frac{1}{2}\log(m_1 + J^{-1}) + \frac{1}{2}\log(m_2 + J^{-1}) - \frac{1}{2}\log(m_1 + m_2 + J^{-1}) - (m_1 + J^{-1})\log(m_1 + J^{-1}) \\ &\quad - (m_2 + J^{-1})\log(m_2 + J^{-1}) + (m_1 + m_2 + J^{-1})\log(m_1 + m_2 + J^{-1}) + J^{-1} \\ &= \frac{J-1}{2}\log(2\pi) + \sum_{j=1}^{J}\Big[\Big(y_1^j + J^{-2} - \frac{1}{2}\Big)\log(y_1^j + J^{-2}) + \Big(y_2^j + J^{-2} - \frac{1}{2}\Big)\log(y_2^j + J^{-2}) \\ &\quad - \Big(y_1^j + y_2^j + J^{-2} - \frac{1}{2}\Big)\log(y_1^j + y_2^j + J^{-2})\Big] + \Big(m_1 + m_2 + J^{-1} - \frac{1}{2}\Big)\log(m_1 + m_2 + J^{-1}) \\ &\quad - \Big(m_1 + J^{-1} - \frac{1}{2}\Big)\log(m_1 + J^{-1}) - \Big(m_2 + J^{-1} - \frac{1}{2}\Big)\log(m_2 + J^{-1}) \end{aligned}$$

Under null hypothesis that Y_1 and Y_2 follow the same distribution, i.e. they share the same probability parameter. Denote the comment probability parameter as $P = (p^1, \cdots, p^J)$. Then

the normal approximation of the multinomial random variables are

$$y_i^j \approx m_i p^j + \sqrt{m_i} z_i^j + \mathcal{O}p(\sqrt{m_i}),\ \textit{for } i = 1, 2;\ j = 1, \cdots, J,$$

where z_i^j's are standard normal random variables and $\sum_{j=1}^{J} z_i^j = 0$ for $i = 1, 2$.

$$\begin{aligned}
LR \approx\ & \frac{J-1}{2}\log(2\pi) + \left(m_1 + m_2 + J^{-1} - \frac{1}{2}\right)\log(m_1 + m_2 + J^{-1}) - \left(m_1 + J^{-1} - \frac{1}{2}\right)\log(m_1 + J^{-1}) \\
& - \left(m_2 + J^{-1} - \frac{1}{2}\right)\log(m_2 + J^{-1}) + \sum_{j=1}^{J}\Bigg[\left(m_1 p^j + \sqrt{m_1} z_1^j + J^{-2} - \frac{1}{2}\right)\log(m_1 p^j + \sqrt{m_1} z_1^j + J^{-2}) \\
& + \left(m_2 p^j + \sqrt{m_2} z_2^j + J^{-2} - \frac{1}{2}\right)\log(m_2 p^j + \sqrt{m_2} z_2^j + J^{-2}) \\
& - \left(m_1 p^j + \sqrt{m_1} z_1^j + m_2 p^j + \sqrt{m_2} z_2^j + J^{-2} - \frac{1}{2}\right)\log(m_1 p^j + \sqrt{m_1} z_1^j + m_2 p^j + \sqrt{m_2} z_2^j + J^{-2})\Bigg] \\
= \ & \frac{J-1}{2}\log(2\pi) + \left(m_1 + m_2 + J^{-1} - \frac{1}{2}\right)\left[\log(m_1 + m_2) + \log\left(1 + \frac{J^{-1}}{m_1 + m_2}\right)\right] \\
& - \left(m_1 + J^{-1} - \frac{1}{2}\right)\left[\log m_1 + \log\left(1 + \frac{J^{-1}}{m_1}\right)\right] - \left(m_2 + J^{-1} - \frac{1}{2}\right)\left[\log m_2 + \log\left(1 + \frac{J^{-1}}{m_2}\right)\right] \\
& + \sum_{j=1}^{J}\Bigg\{\left(m_1 p^j + \sqrt{m_1} z_1^j + J^{-2} - \frac{1}{2}\right)\left[\log(m_1 p^j) + \log\left(1 + \frac{\sqrt{m_1} z_1^j + J^{-2}}{m_1 p^j}\right)\right] \\
& + \left(m_2 p^j + \sqrt{m_2} z_2^j + J^{-2} - \frac{1}{2}\right)\left[\log(m_2 p^j) + \log\left(1 + \frac{\sqrt{m_2} z_2^j + J^{-2}}{m_2 p^j}\right)\right] \\
& - \left((m_1 + m_2)p^j + \sqrt{m_1} z_1^j + \sqrt{m_2} z_2^j + J^{-2} - \frac{1}{2}\right)\Big[\log((m_1 + m_2)p^j) \\
& + \log\left(1 + \frac{\sqrt{m_1} z_1^j + \sqrt{m_2} z_2^j + J^{-2}}{(m_1 + m_2)p^j}\right)\Big]\Bigg\} \\
= \ & \frac{J-1}{2}\log(2\pi) + \frac{J-1}{2}\log\left(\frac{1}{m_1} + \frac{1}{m_2}\right) + \left(m_1 + m_2 + J^{-1} - \frac{1}{2}\right)\log\left(1 + \frac{J^{-1}}{m_1 + m_2}\right) \\
& - \left(m_1 + J^{-1} - \frac{1}{2}\right)\log\left(1 + \frac{J^{-1}}{m_1}\right) - \left(m_2 + J^{-1} - \frac{1}{2}\right)\log\left(1 + \frac{J^{-1}}{m_2}\right) \\
& + \sum_{j=1}^{J}\Bigg\{\left(m_1 p^j + \sqrt{m_1} z_1^j + J^{-2} - \frac{1}{2}\right)\log\left(1 + \frac{\sqrt{m_1} z_1^j + J^{-2}}{m_1 p^j}\right) \\
& + \left(m_2 p^j + \sqrt{m_2} z_2^j + J^{-2} - \frac{1}{2}\right)\log\left(1 + \frac{\sqrt{m_2} z_2^j + J^{-2}}{m_2 p^j}\right) \\
& - \left((m_1 + m_2)p^j + \sqrt{m_1} z_1^j + \sqrt{m_2} z_2^j + J^{-2} - \frac{1}{2}\right)\log\left(1 + \frac{\sqrt{m_1} z_1^j + \sqrt{m_2} z_2^j + J^{-2}}{(m_1 + m_2)p^j}\right)\Bigg\}
\end{aligned}$$

Note that, in general, by L'Hopital's rule, as $m_i \to \infty$,

$$\sqrt{m_i}\log\left(1 + \frac{\sqrt{m_i} z_i^j + J^{-2}}{m_i p^j}\right) = \frac{\log\left(1 + \frac{\sqrt{m_i} z_i^j + J^{-2}}{m_i p^j}\right)}{1/\sqrt{m_i}} \longrightarrow \frac{z_i^j}{p^j},$$

for $i = 1, 2;\ j = 1, \cdots, J$.

Under the assumption that m_1 and m_2 are increasing at the same rate, let $m_1 = m$ and

$m_2 = cm$, for some $c > 0$. Then as $m \to \infty$,

$$\begin{aligned} & (\sqrt{m_1} z_1^j + \sqrt{m_2} z_2^j) \log\left(1 + \frac{\sqrt{m_1} z_1^j + \sqrt{m_2} z_2^j + J^{-2}}{(m_1 + m_2) p^j}\right) \\ = \quad & (z_1^j + \sqrt{c} z_2^j)\sqrt{m} \log\left(1 + \frac{\sqrt{m}(z_1^j + \sqrt{c} z_2^j) + J^{-2}}{(1+c) m p^j}\right) \\ \longrightarrow \quad & \frac{(z_1^j + \sqrt{c} z_2^j)^2}{(1+c)p^j}, \ \textit{for } j = 1, \cdots, J. \end{aligned}$$

Therefore, as $m \to \infty$, the log likelihood ratio

$$\begin{aligned} & LR \\ \approx \quad & \frac{J-1}{2}\log(2\pi) + \frac{J-1}{2}\log\frac{1+c}{cm} + \left(m(1+c) + J^{-1} - \frac{1}{2}\right)\log\left(1 + \frac{J^{-1}}{m(1+c)}\right) \\ & - \left(m + J^{-1} - \frac{1}{2}\right)\log\left(1 + \frac{J^{-1}}{m}\right) - \left(cm + J^{-1} - \frac{1}{2}\right)\log\left(1 + \frac{J^{-1}}{cm}\right) \\ & + \sum_{j=1}^{J} \Bigg\{ \left(mp^j + \sqrt{m} z_1^j + J^{-2} - \frac{1}{2}\right)\log\left(1 + \frac{\sqrt{m} z_1^j + J^{-2}}{mp^j}\right) \\ & + \left(cmp^j + \sqrt{cm} z_2^j + J^{-2} - \frac{1}{2}\right)\log\left(1 + \frac{\sqrt{cm} z_2^j + 1/25}{cmp^j}\right) \\ & - \left((1+c)mp^j + \sqrt{m} z_1^j + \sqrt{cm} z_2^j + J^{-2} - \frac{1}{2}\right)\log\left(1 + \frac{\sqrt{m} z_1^j + \sqrt{cm} z_2^j + J^{-2}}{(1+c)mp^j}\right)\Bigg\} \\ \longrightarrow \quad & -\infty \end{aligned}$$

Therefore, Y_1 and Y_2 have the same cluster label almost surely.

1.4 *Determine the threshold d*

The size of S_2^d is a step function of d. We suggest to plot the size of S_2^d against a decreasing series of cutoffs. We approximate the curvature of the plot by looking at the total segment length of every consecutive Δ number of steps. Then pick the point whose left Δ steps minus its right Δ steps is the largest as the optimal point. The default Δ value is 3 in our program. Larger Δ values lead to more coarse yet more robust approximation of the curvature. We also require a minimum length for the step on the left of the optimal point to guarantee that

the noise set did not enlarge shortly after the value that is slightly greater than the cutoff. If the step on the left of the optimal point is shorter than the required minimum length, we move the optimal point to the left by one step and the check the length of the next step. The final threshold is chosen to be the optimal point shifted to the left by the minimum length. This default sets the minimum length to be half ($\alpha = 0.5$) of the average length of the left Δ steps. A larger α leads to a more conservative result while a smaller α corresponds to a more liberal result. Both Δ and $\alpha = 0.5$ are introduced to mathematically capture the boundary of noise and signal sections of the data. In practice, we suggest users to verify the output by examining the site count plot directly.

1.5 *Experimental setup flow charts*

[Figure 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

1.6 *Simulated data*

Consider the following experiment setup for a viral population with genome length 300 nucleotides and five possible nucelotides at each genome site: A, C, G, T, M. The simulation mimics the experiment which first samples the RNA data twice before the administration of the treatment (t1, t2), then obtains a control group (t3) and a treatment group (t3D). For each genome site at time t_1, t_2, t_3, the sequencing read count data are generated from multinomial distribution invariant in time and dependent on the genomic location. For the treated group t_{3D}, the evolved drug resistance sites 1, 21, 41, 61, 81 are generated from alternative multinomial distributions, while the rest follow the same multinomial distributions as other time points.

As discussed in the model description, we assume that the probability parameter for the

nucleotide at each genomic location is sampled from a Dirichlet mixture model. For the sample without treatment, $P_1, \cdots, P_{15}$, were used to generate the five possible reads: A, C, G, T, M. Five additional probability parameters, $P_{16}, \cdots, P_{20}$, were introduced to generate the treated population. The total number of mixture component for the joint dataset is therefore 20.

At each genomic location i, the corresponding summary statistics are

$$\begin{aligned} Ht(D_i) &= \min\{Ht(\pi_i^{t_1}, \pi_i^{t_{3D}}), Ht(\pi_i^{t_2}, \pi_i^{t_{3D}})\} \\ Ht(N_i) &= \max\{Ht(\pi_i^{t_1}, \pi_i^{t_2}), Ht(\pi_i^{t_1}, \pi_i^{t_3}), Ht(\pi_i^{t_2}, \pi_i^{t_3})\} \end{aligned} \tag{4}$$

1.7 *Additional IVA Ht plot*

In this subsection, we provide the Ht plots for segment 8 using five time points.

[Figure 4 about here.]

[Figure 5 about here.]

1.8 *Analysis for IVA with four passages*

For a fairer comparison to Foll et al, we applied our method to the joint data from Passages 1, 3, 9, 12 for both the control and treatment groups, i.e. $t_1, t_2, t_3, t_4, t_{3D}, t_{4D}$. The summary statistics used are

$$\begin{aligned} Ht(D_i) &= \min\{Ht(\pi_i^{t_1}, \pi_i^{t_{4D}}), Ht(\pi_i^{t_2}, \pi_i^{t_{4D}})\} \\ Ht(N_i) &= \max\{Ht(\pi_i^{t_1}, \pi_i^{t_2}), Ht(\pi_i^{t_1}, \pi_i^{t_j}), Ht(\pi_i^{t_2}, \pi_i^{t_j}),\ j = 3, 4\} \end{aligned} \tag{5}$$

The result from each biological replicate is shown in Table 2. Here we used the default parameters, $\Delta = 3,\ \alpha = 0.5$, and the summary results are for without the *Gibbs* step.

[Table 2 about here.]

Taking the intersection of the findings from both replicates, we identify only S6-822 as a substitution site due to the treatment (although potentially S2-32, if the S2-31 and S2-32 sites of our analysis are treated as one). Several other sites, S1-2299, S2-2303, S3-173,

174,176,200, 203, 2078, 2192, 2193, 2195, S5-24, 389,1103, S6-977, were identified as locations with evolutionary changes not due to the treatment.

As for Foll et al, the following sites are identified drug resistant: S2-32, S3-2193, S4-47, S4-1394, S6-581, S6-822, S7-146, S8-819; the sites with evolutionary changes without treatment are S2-1118, S4-1394, S5-1103, S5-1395.

1.9 *Raw data plot*

Nucleotide read count proportion plot for identified IVA sites. The color tiles on the top of each panel indicates the total read count at each time point. The top and bottom rows are for each replicate; the left and right panels are for control and treatment groups, respectively.

[Figure 6 about here.]

The nucleotide read type proportions at each time point for the sites with high genetic variation that might be due to adaptation to the hosts: S1-2299, S3-2193, S4-1210, S5-24, S5-1103, S7-91, S8-819.

[Figure 7 about here.]

[Figure 8 about here.]

[Figure 9 about here.]

[Figure 10 about here.]

[Figure 11 about here.]

[Figure 12 about here.]

[Figure 13 about here.]

[Figure 14 about here.]

[Figure 15 about here.]

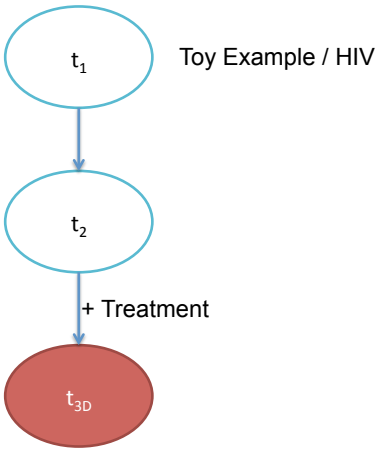

Figure 1. Illustration of the experimental setup the toy example and the HIV data. This setup includes two untreated populations (t_1, t_2) and one post-treatment population (t_{3D}). Observations are collected from each time point.

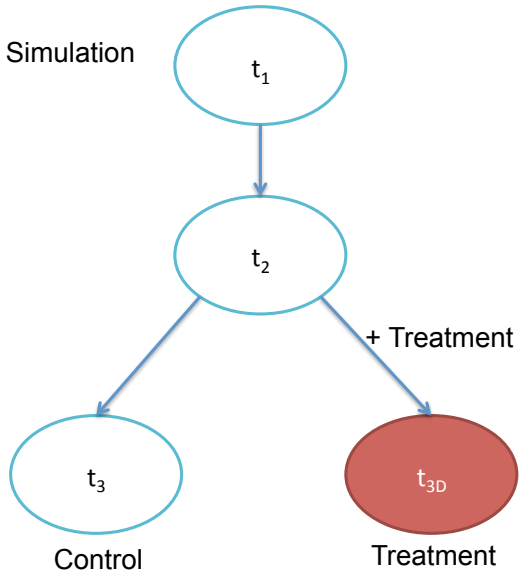

Figure 2. The experimental design used for the simulated test data. After two generations without treatment (t_1, t_2) the population is split into a control group t_3 and a treatment group t_{3D}. The treatment group is given a drug and allowed to evolve resistance to that drug for a few generations. The before treatment, control group, and treatment populations are sampled and sequenced.

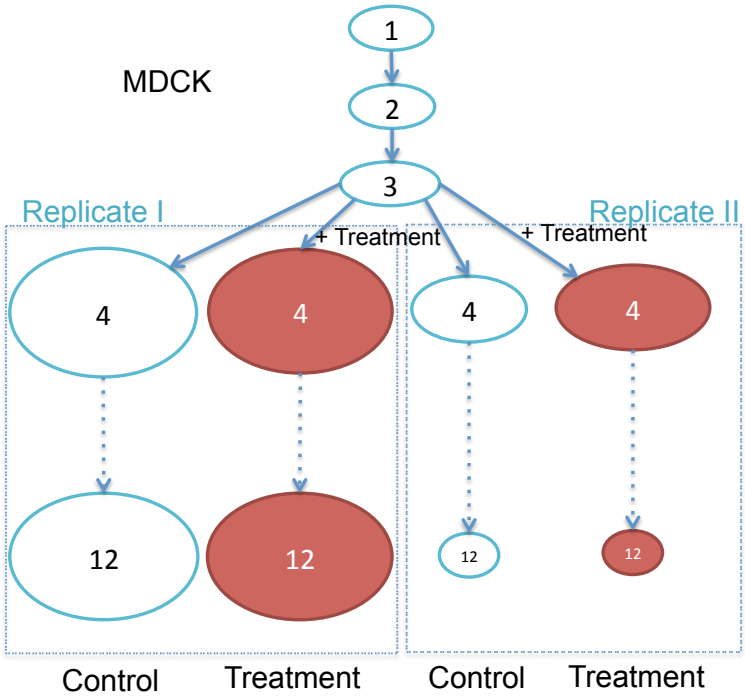

Figure 3. Only the first 12 passages were used in Foll et al (**?**). The complete dataset includes two biological replicates with one control group and one treatment group. Each ovals presents a passage. The colors white and red indicate absence and presence of the inhibitor. The sizes of the ovals indicate the average total read count per genome site. Note that the first replicate have much larger total reads than the second.

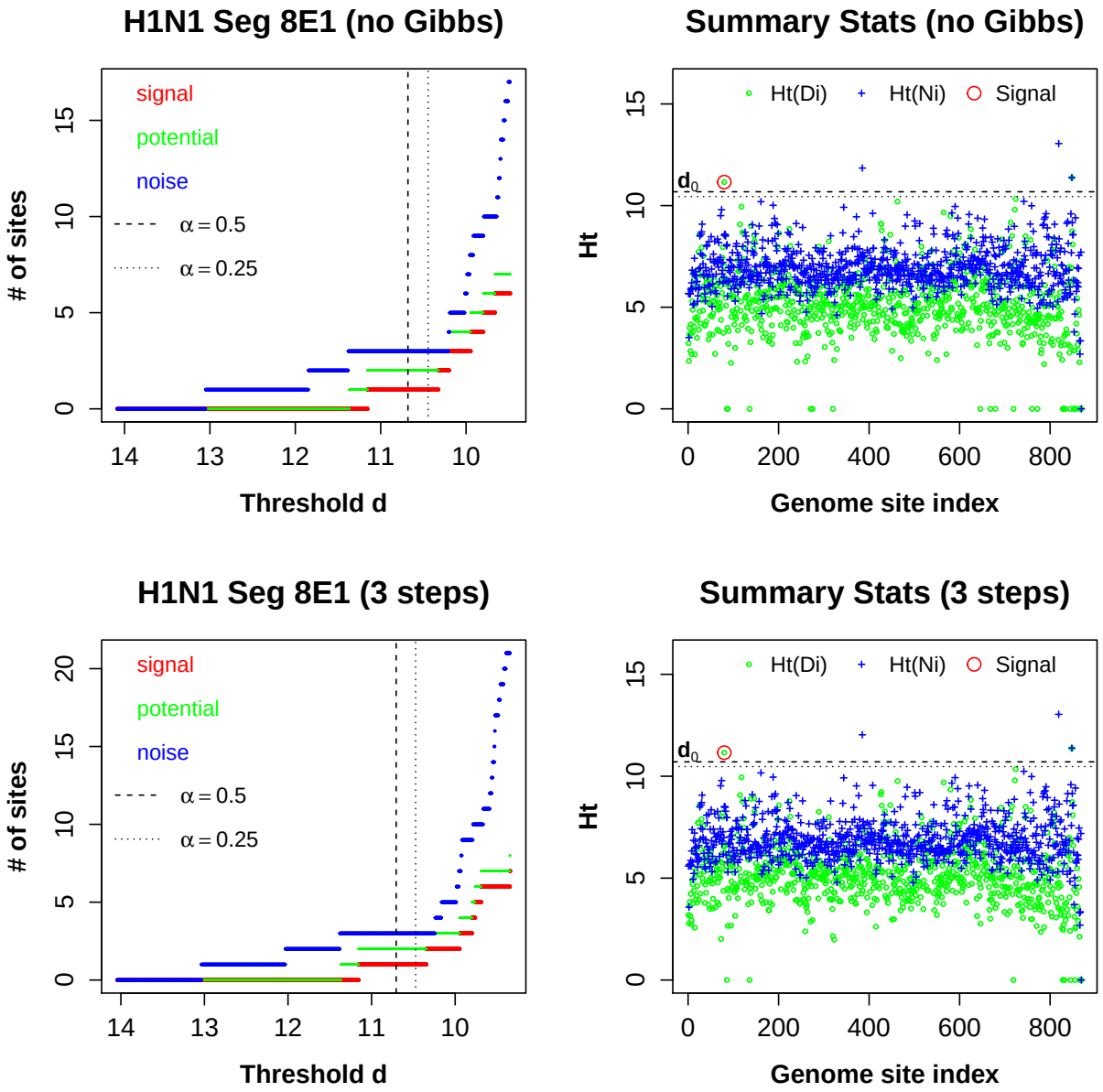

Figure 4. The result plots for H1N1 Seg8E1. The inference result for H1N1 Seg7E2 is consistent even without the *Gibbs* step, and is robust to the choices of α parameter. The highlighted site, S8-80, was identified as a drug resistant site with or without the *Gibbs* step. The noise set consists of S8-385, S8-819, S8-848.

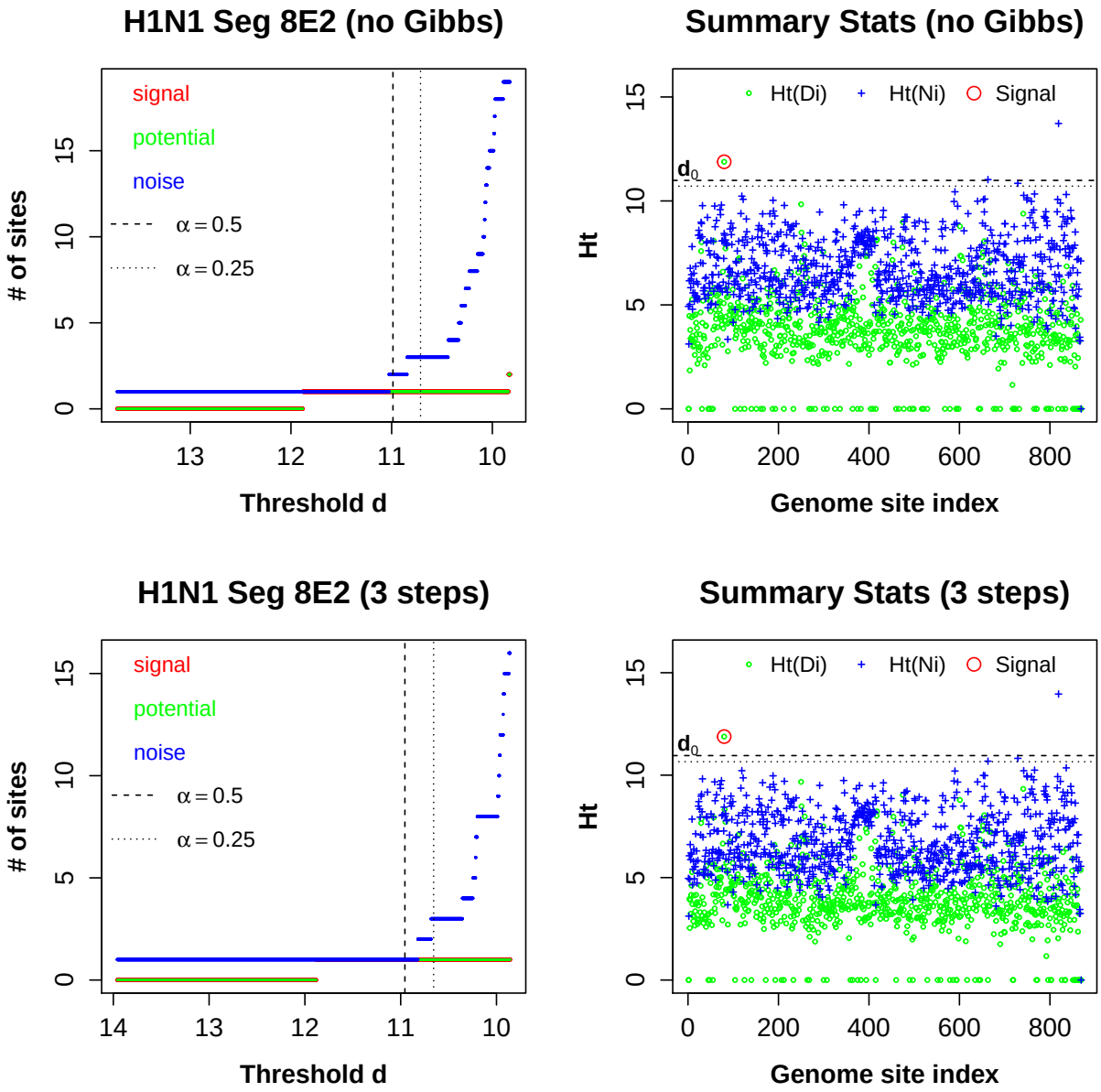

Figure 5. Similar to Seg8E1, the inference result is consistent with or without the *Gibbs* step. The highlighted S8-80 was identified as a signal site; S8-819 was a noise site that was also identified in Seg8E1.

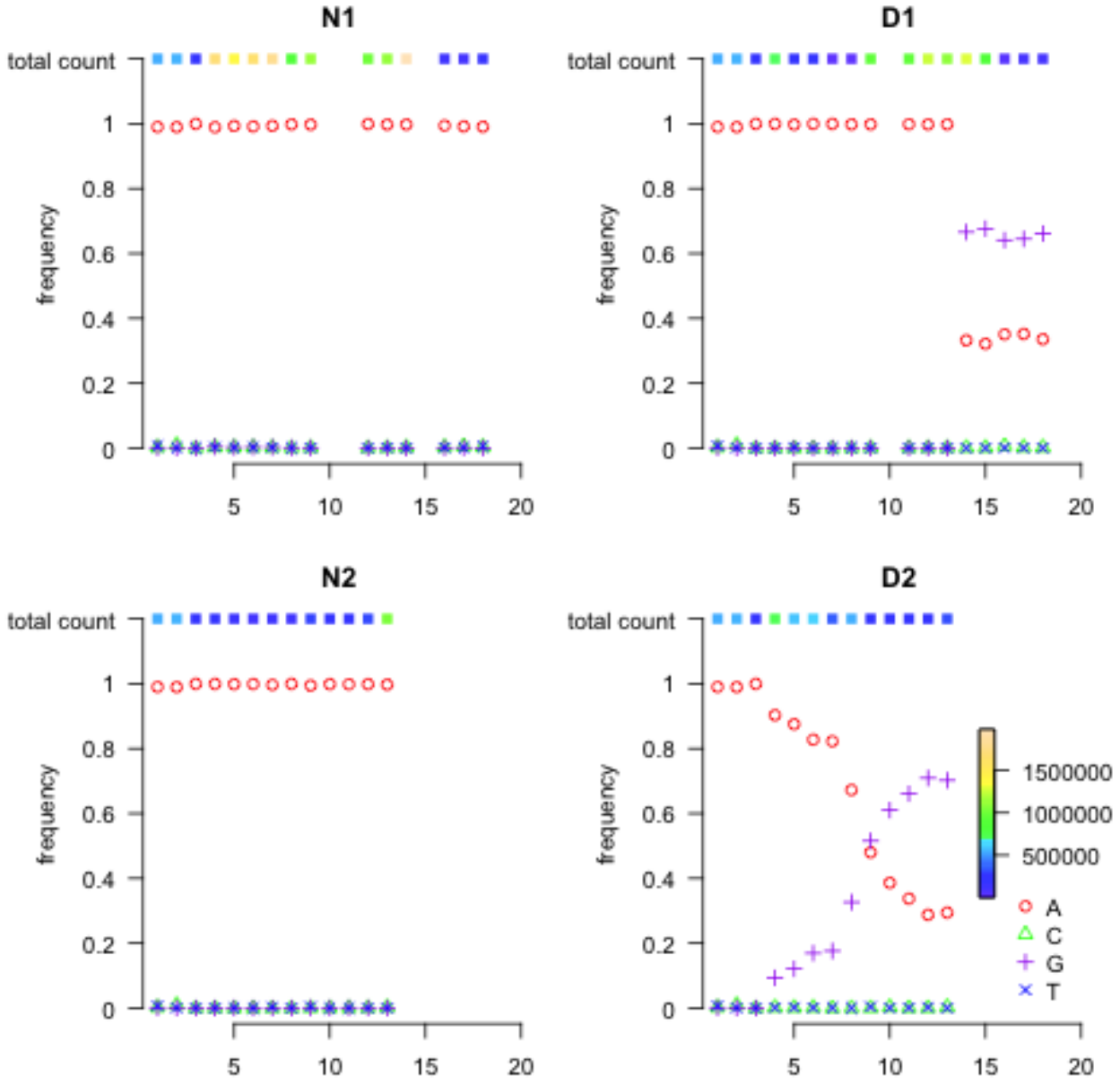

Figure 6. H1N1 nucleotide read count proportion and total count at position S8-80. Unlike the previous two figures, the read type does not switch completely. Instead, the starting read type A remains in more than 25% of the sample while the rest have read type G post treatment.

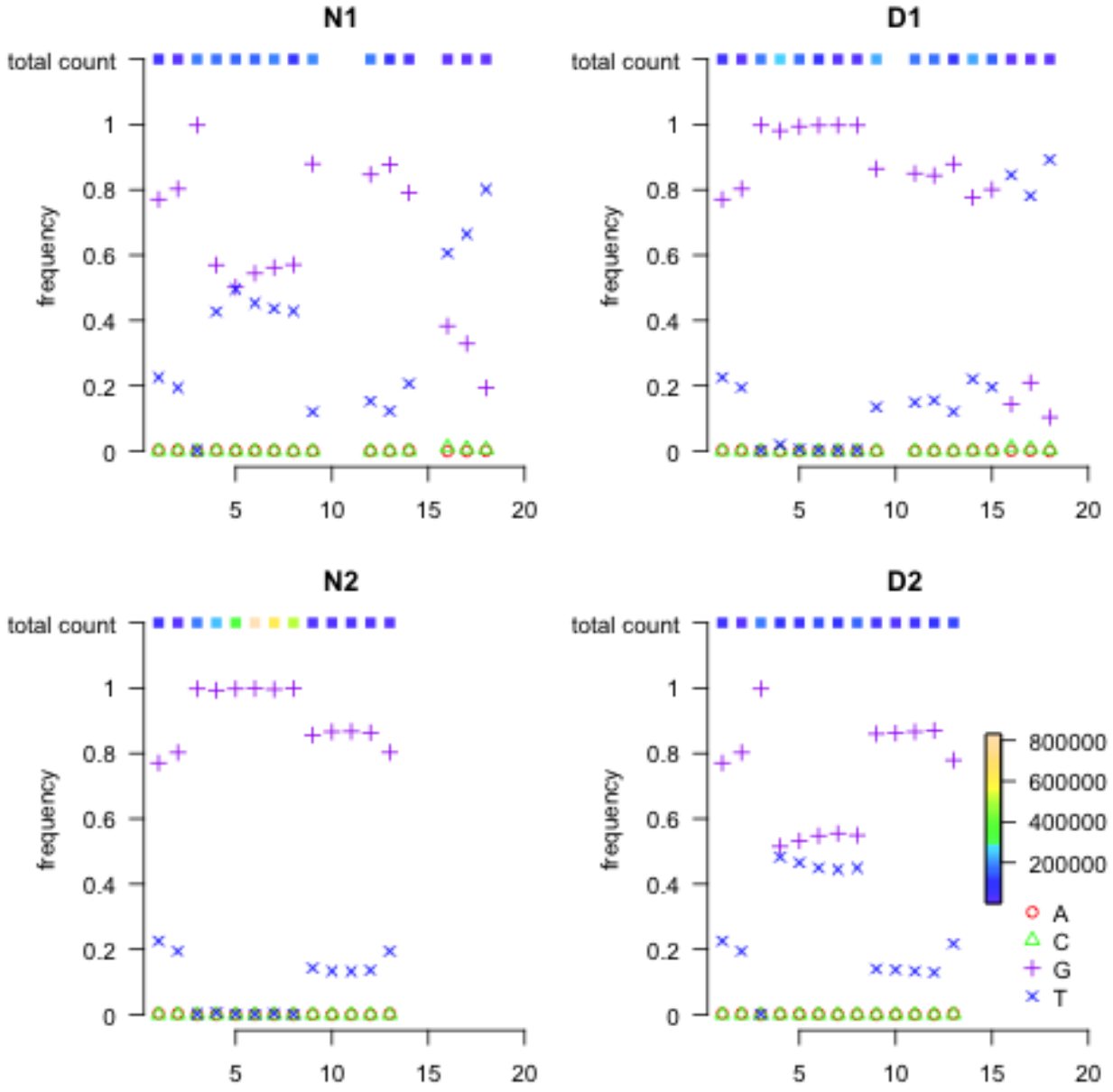

Figure 7. H1N1 nucleotide read count proportion and total count at position S1-2299. In all four panels, there is great variation between read type G and T. The alternating behavior happened in almost all panels in dictating that the genetic variation is not due to treatment and likely that the read types G and T together dominate the reads at equilibrium.

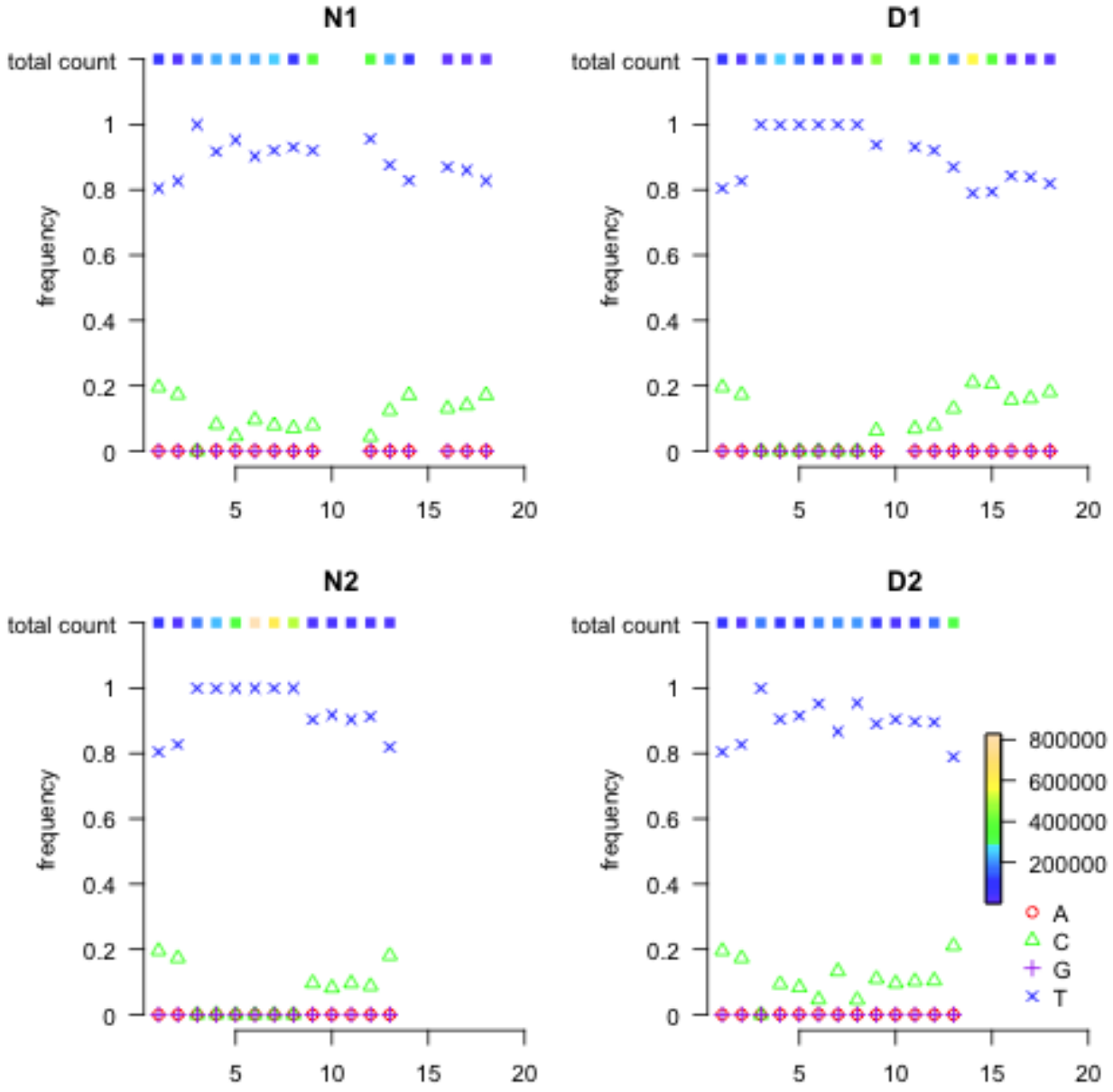

Figure 8. H1N1 nucleotide read count proportion and total count at position S1-2303. Regardless treatment or control, there was significant fluctuation in the mixture proportion of nucleotides C and T over time.

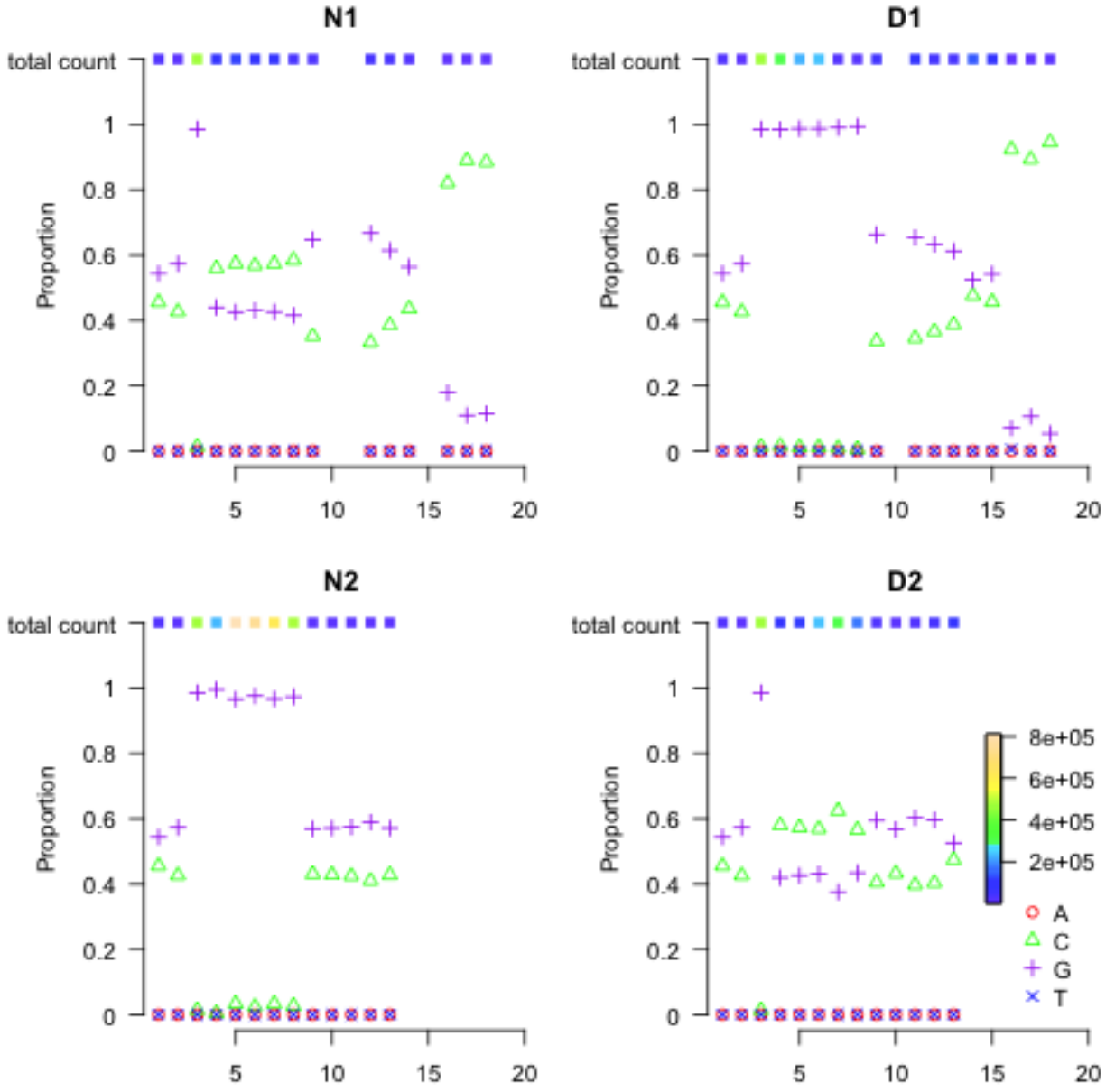

Figure 9. H1N1 nucleotide read count proportion and total count at position S3-2193. In all four panels, there is great variation between read type C and G with and without the treatment.

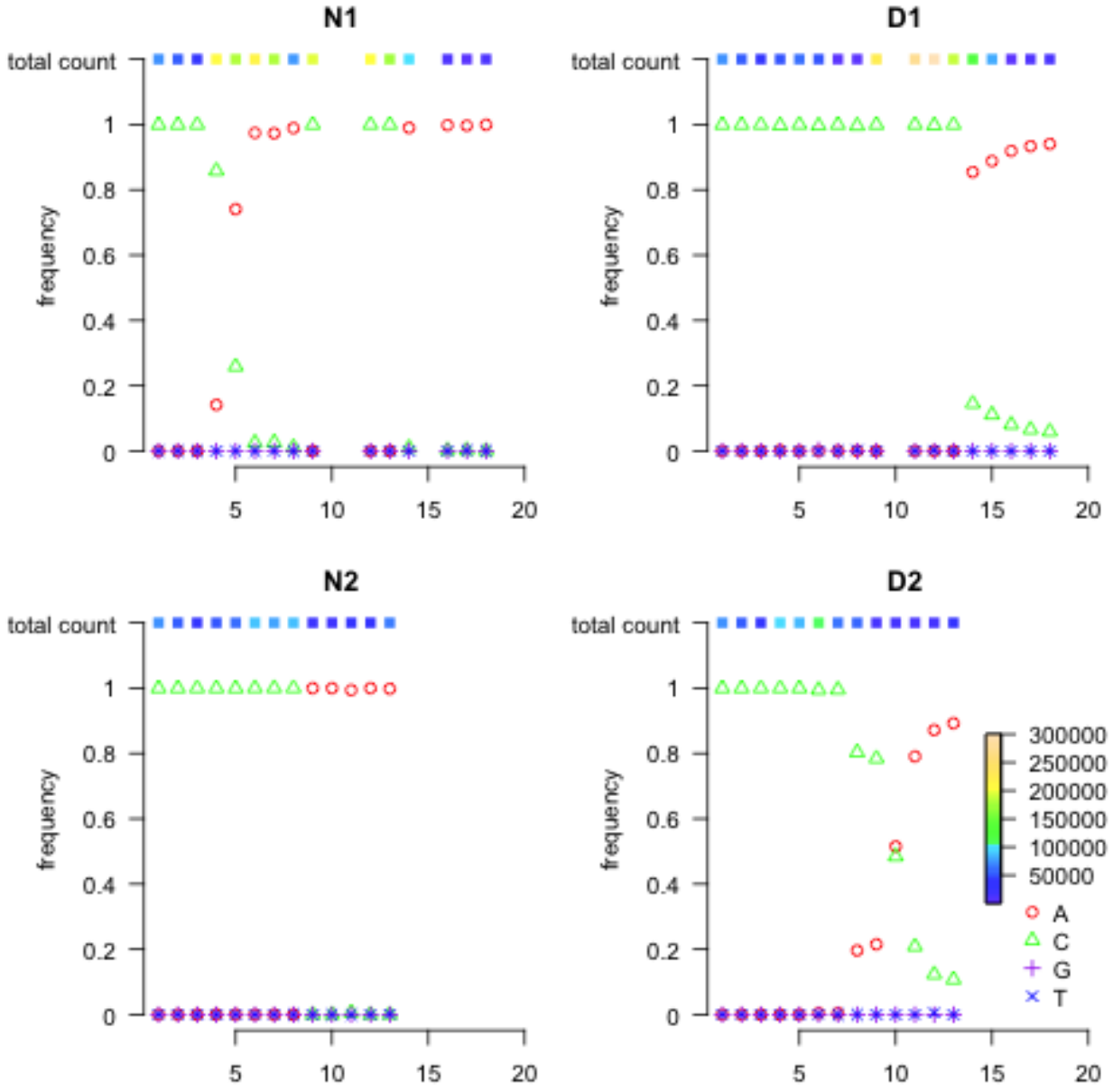

Figure 10. H1N1 nucleotide read count proportion and total count at position S4-1210. There is a complete transversion from C to A in all four panels, suggesting that this change might be due to adaptation to the hosts or genetic drifts.

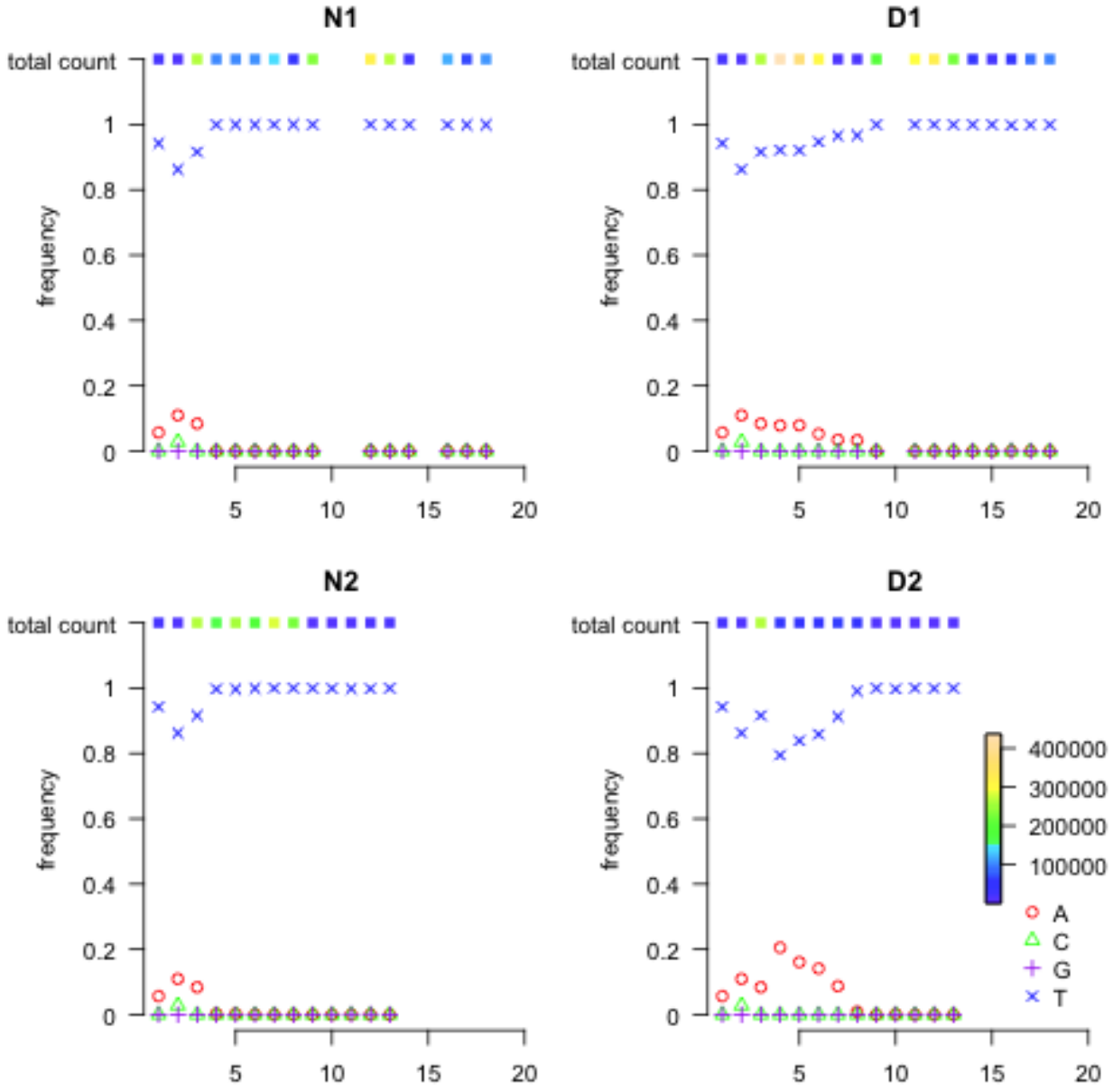

Figure 11. H1N1 nucleotide read count proportion and total count at position S5-24. The four panels show similar pattern. It appears that the treated groups (right two panels) took longer time to reach equilibrium.

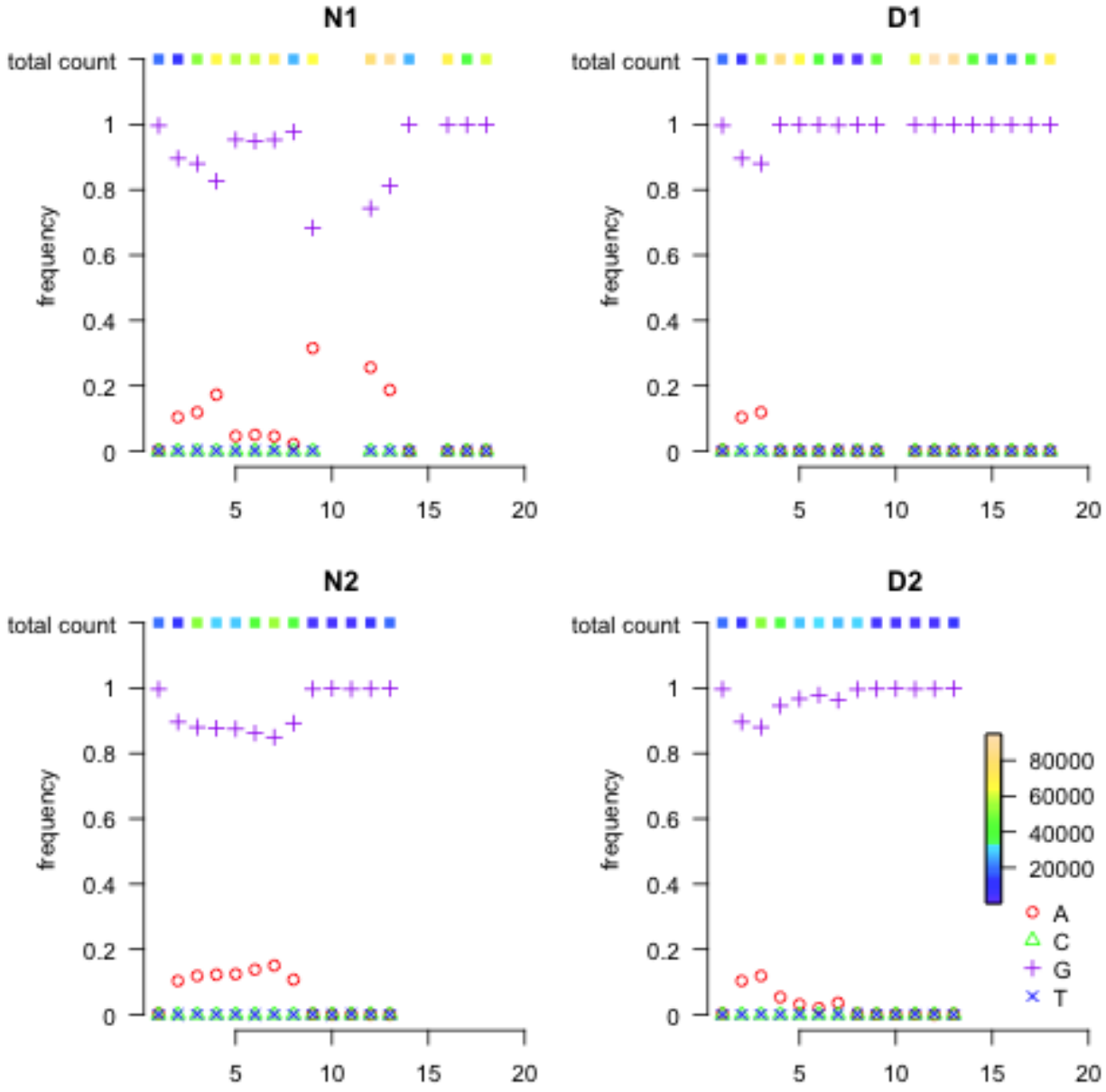

Figure 12. H1N1 nucleotide read count proportion and total count at position S5-389. The untreated groups appear to present greater variation over time.

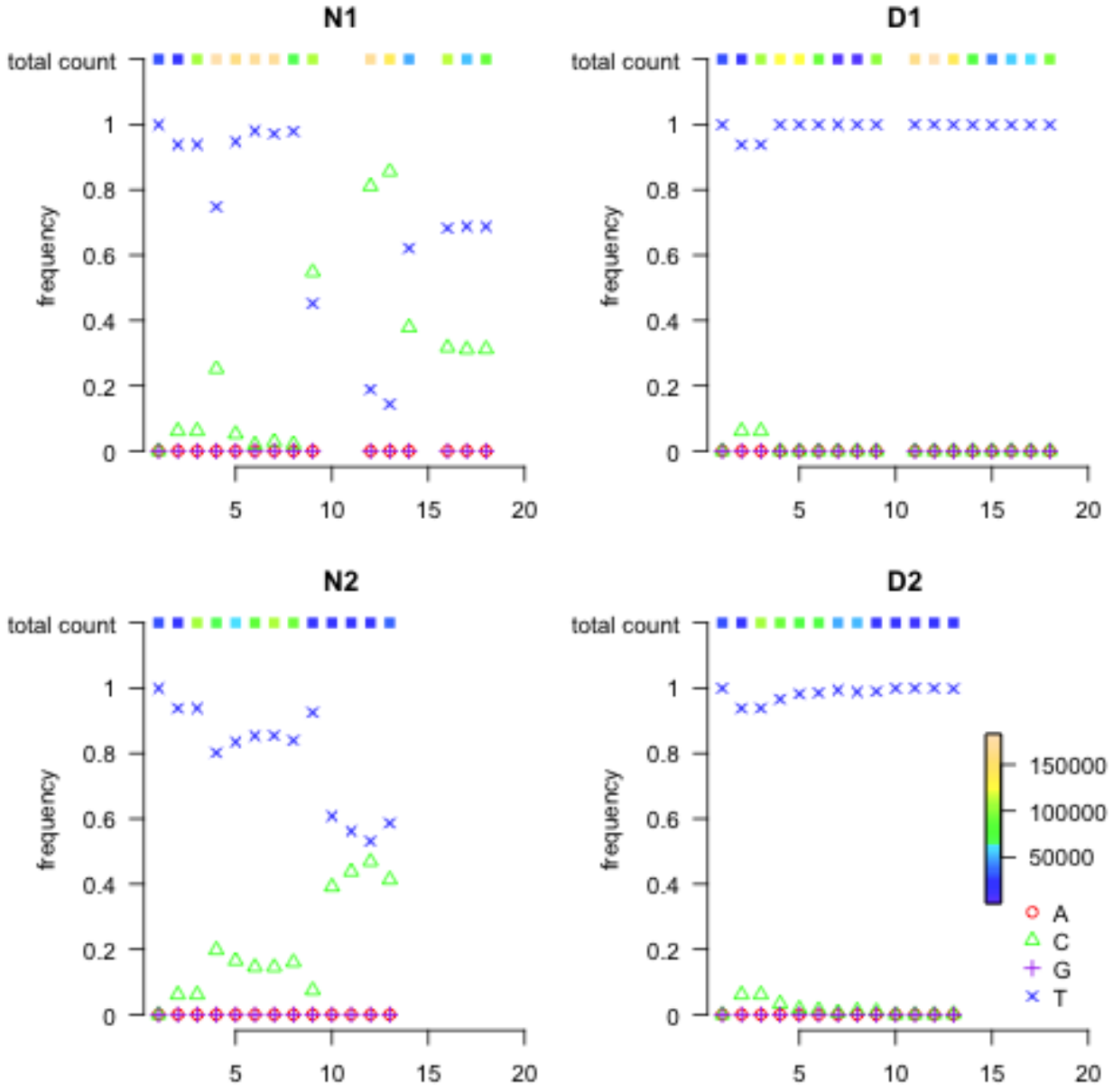

Figure 13. H1N1 nucleotide read count proportion and total count at position S5-1103. The untreated groups appear to present greater variation over time.

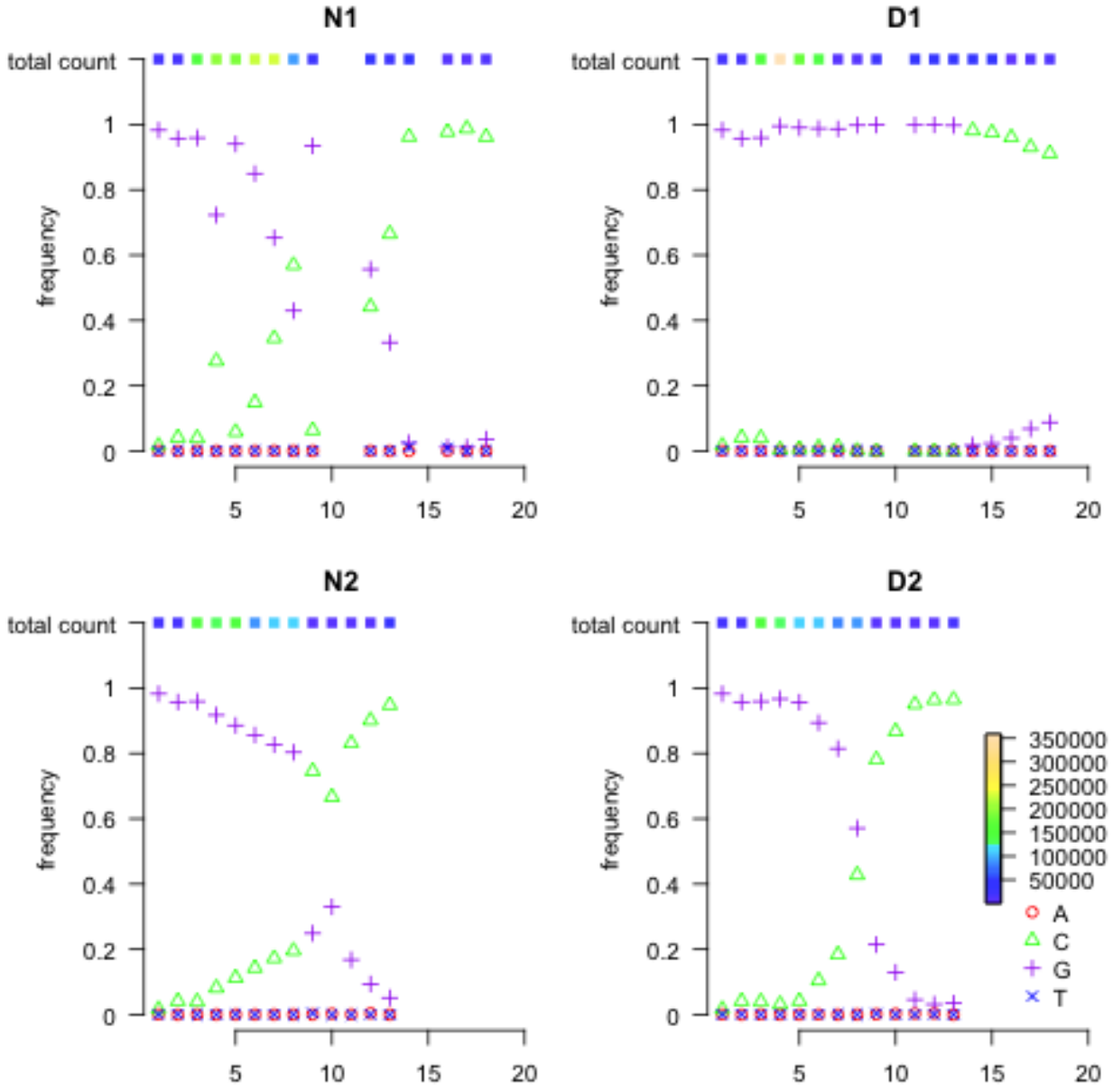

Figure 14. H1N1 nucleotide read count proportion and total count at position S7-91. All four panels show complete transversion from G to C.

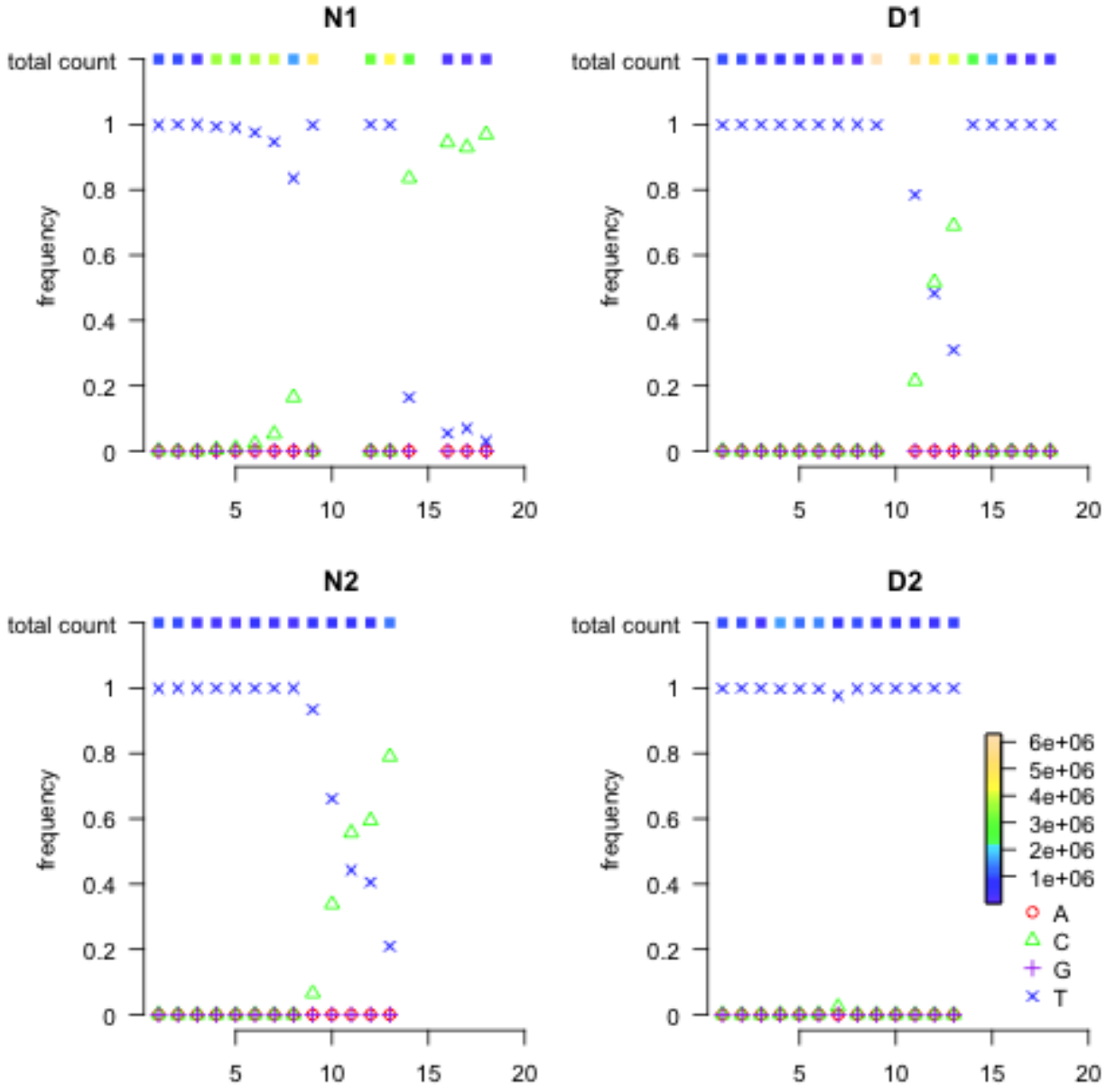

Figure 15. H1N1 nucleotide read count proportion and total count at position S8-819. The untreated groups present transition from G to C; one of the treated groups (top right panel) appears to have interchanged read type in the intermediate time points while the other treated group presents low variation.

	$Y_1^{t_1}$	$Y_2^{t_1}$	$Y_3^{t_1}$	$Y_4^{t_1}$	$Y_5^{t_1}$
A	8	0	0	0	0
C	0	0	6	1	0
G	0	0	0	0	0
T	0	7	0	6	8
M	0	0	0	0	0

	$Y_1^{t_2}$	$Y_2^{t_2}$	$Y_3^{t_2}$	$Y_4^{t_2}$	$Y_5^{t_2}$
A	10	0	0	0	0
C	0	0	9	2	0
G	0	0	0	0	0
T	0	10	0	9	5
M	0	0	0	0	0

	$Y_1^{t_3}$	$Y_2^{t_3}$	$Y_3^{t_3}$	$Y_4^{t_3}$	$Y_5^{t_3}$
A	11	0	0	0	0
C	0	0	10	0	0
G	0	0	0	0	0
T	3	9	0	7	8
M	0	0	1	0	1

$\Downarrow$

	$Y_1^{all\ t}$	$Y_2^{all\ t}$	$Y_3^{all\ t}$	$Y_4^{all\ t}$	$Y_5^{all\ t}$	$Y_6^{all\ t}$	$Y_7^{all\ t}$	$Y_8^{all\ t}$
A	18	0	0	0	0	11	0	0
C	0	15	0	1	2	0	10	0
G	0	0	0	0	0	0	0	0
T	0	0	46	6	9	3	0	8
M	0	0	0	0	0	0	1	1

$\Downarrow$

$$
\begin{aligned}
Y_1^{all\ t} &= Y_1^{t_1} + Y_1^{t_2},\\
Y_2^{all\ t} &= Y_3^{t_1} + Y_3^{t_2},\\
Y_3^{all\ t} &= Y_2^{t_1} + Y_5^{t_1} + Y_2^{t_2} + Y_5^{t_2} + Y_2^{t_3} + Y_4^{t_3} + Y_5^{t_3},\\
Y_4^{all\ t} &= Y_4^{t_1},\\
Y_5^{all\ t} &= Y_4^{t_2},\\
Y_6^{all\ t} &= Y_1^{t_3},\\
Y_7^{all\ t} &= Y_3^{t_3},\\
Y_8^{all\ t} &= Y_5^{t_3}.
\end{aligned}
$$

Table 1

Toy example of joining and preprocessing three 5×5 data matrices. The first few columns in the joint data matrix (second row) are the consolidation of columns with single nucleotide read in the sampled data panels (first row). The remaining columns of the joint data matrix are the copies of non-homogeneous reads of the sample (first row). The detail of the consolidation process is described in the panel in the third row.

Seg	E	d_0	$S_3^{d_0}$	$S_2^{d_0}$
1	1	8.713	$\emptyset$	33, 281, 404, 824, 839, 926, 1889, 2290, 2298, 2299, 2303
	2	9.138	2072	311, 2299, 2303
2	1	9.718	32	224, 1118, 1499, 2066
	2	9.43	31, 1663, 2099	1483, 1675, 2033
3	1	7.84	1613	89, 134, 173, 174, 176, 177, 200, 203, 1556, 2078, 2192, 2193, 2195
	2	7.944	180, 997	79, 146, 153, 173, 174, 176, 200, 203, 210, 1527, 1850, 2078, 2192, 2193, 2195
4	1	9.447	47, 1394	729
	2	9.274	$\emptyset$	638, 1210
5	1	8.855	$\emptyset$	24, 389, 1103, 1395
	2	8.522	300	24, 389, 1103
6	1	8.496	581, 822	977,1047
	2	7.728	822	680, 977
7	1	9.099	146, 1005	$\emptyset$
	2	8.725	91, 1004	637
8	1	10.307	200, 819	385
	2	10.435	80	729, 819

Table 2

Our approach (w/o Gibbs) identifies only one true signal when data from only Passages 1, 3, 9, and 12 are used. The thresholds and corresponding signal & noise sets for each segment according to each biological replicate. The sites identified as signal in both experiments are highlighted in red, the ones identified as noise in both experiments are highlighted in blue. Fewer substitution sites were identified compared to previous table (see Table **??***).*