How do antigenically varying pathogens avoid cross-reactive responses to invariant antigens?

Philip L. F. Johnson¹,* , Beth F. Kochin¹, Rafi Ahmed² and Rustom Antia¹,*

Pathogens such as trypanosomes and malaria use antigenic variation to evade immune responses and prolong the duration of infections. As pathogens typically express more than one antigen, even relatively rare conserved antigens might be expected to trigger cross-reactive immune responses capable of clearing the infection. We use simple mathematical models that explicitly consider the dynamic interplay between the replicating pathogen, immune responses to different antigens and immune exhaustion to explore how pathogens can escape the responses to both variable and invariant (conserved) antigens. Our results suggest two hypotheses. In the first, limited quantities of invariant antigens on each pathogen may lead to saturation in killing by cross-reactive responses. In the second, antigenic variation of the dominant antigen prolongs the duration of infection sufficiently to allow for exhaustion of the cross-reactive responses to subdominant, invariant epitopes prior to their being able to control the infection. These hypotheses make distinct predictions: the former predicts that cross-reactive responses will always be ineffective while the latter predicts that appropriately timed treatment could, by preventing exhaustion, lead to the generation of long-lasting protective cross-reactive immunity and thus act similarly to a vaccine.

Keywords: immunodominance; immune exhaustion; mathematical model; chronic infection

1. INTRODUCTION

The adaptive immune system is one of the primary factors that limits the replication of pathogens within a host. As adaptive immune responses target specific antigens expressed by the pathogen, some pathogens evade these responses by changing their antigens. These changes can occur either within a host during the course of a single infection (‘antigenic variation’) or at the epidemiological level (‘antigenic diversity’). We focus on pathogens such as Trypanosoma brucei, Trypanosoma cruzi and the plasmodia responsible for malaria for which antigenic variation is generally assumed to be responsible for persistent infection [1,2].

Explaining how antigenic variation enables pathogens to persist for many months [1,3] becomes more difficult when we realize that pathogens express multiple antigens simultaneously. As shown in figure 1, changing a single antigen allows a new pathogen variant to escape immunity to the previous version of that particular antigen, but this new variant will still be susceptible to the cross-reactive immune responses directed against unchanged antigens. Clearly, the magnitude of the benefit accruing from variation is greatest if a pathogen changes its immunodominant antigen—that is, the antigen that elicits the strongest immune response.

We perform simple calculations to determine how the duration of an infection depends on the characteristics of the variable and invariant (conserved) antigens, and the immune responses that they elicit. Our calculations suggest that even poorly immunogenic conserved antigens might be expected to elicit cross-reactive responses sufficient to control the pathogen on a relatively short timescale. Thus, we are left with a puzzle: why do cross-reactive responses not prevent long-lasting ‘chronic’ infections?

We use models to explore how this apparent paradox might be explained by considering how the assumptions in our initial calculations may need to be revised in light of recent research.

The first assumption we address is the standard ‘mass action’ term for killing of pathogens by the immune responses. This term is proportional to the product of the densities of pathogen and immune response, which may be reasonable when the density of an antigen on the pathogen is relatively high, as should be the case for immunodominant antigens. However, more detailed stoichiometric models [4,5] are needed when the density of antigen on the pathogen is very low, as might be the case for a subdominant invariant antigen. For example, with antibody-mediated killing, we might expect that once the few invariant sites are occupied by antibodies, further increases in antibody concentration will not lead to more rapid clearance or killing.

The second assumption we address involves how the immune response grows and is maintained. The theory of clonal selection proposes that pathogens stimulate the clonal expansion of antigen-specific T and B cells to generate large populations that can control the infection. However, experimental findings have shown that this process breaks down during chronic infections when immune
cells are exposed to specific antigen for extended periods of time. Persistent stimulation causes the corresponding immune cells to become dysfunctional and may even result in their death [6]. This phenomenon, termed immune exhaustion, has been documented for both T cell-mediated [7–9] and antibody responses [10,11]. Relatively, few models of immune system dynamics have included exhaustion [12–14], and none of these have also considered antigenically varying pathogens. The interplay between exhaustion and variation becomes particularly intriguing in light of a very recent report showing T cell exhaustion during malaria infections [15].

Our work builds on a rich history of models used to study the dynamics of antigenic variation [16–27]. We elaborate on the connections to this earlier work in the Discussion, where we make predictions for treatment and vaccination that distinguish our hypotheses from those of earlier studies.

2. MODEL AND RESULTS
For simplicity, we describe our models using ordinary differential equations. We simulate stochastic trajectories from these equations in order to capture the stochastic effects inherent in antigenic variation by using a Gillespie-style algorithm. Specifically, we perform these simulations using the adaptive tau leaping algorithm [28] run from R [29] as implemented in the ‘adaptivetau’ R package, which is freely available at http://cran.r-project.org/web/packages/adaptivetau/index.html.

Rather than developing complex models in an attempt to capture all known details of immune system dynamics, we will make simple models with the goal of discovering fundamental principles that could explain the dynamics of antigenically varying pathogens [30]. Indeed, given the limitations in our quantitative understanding of the detailed dynamics of immune responses and uncertainties in many of the parameters of the models, simpler models frequently provide more robust results than complex models [31,32]. While these uncertainties make it hard to parameterize the model for a particular infection, we use biologically reasonable parameter ranges and discuss how our model might apply to trypanosome and malaria infections.

(a) Basic model
Consider the following model for the dynamics of an antigenically varying pathogen and the adaptive immune responses it elicits. Let $P_i$ be the density of variant $i$, $X_i$ be the level of the corresponding variant-specific immune response and $Z$ be the level of the cross-reactive immune responses. The pathogen is assumed to exhibit logistic growth in the absence of immunity. The immune response (either B or T cell) is assumed to grow by clonal expansion of antigen-specific cells in a manner dependent on the amount of antigen/pathogen. Putting these factors together, we have:

$$
\frac{dP_i}{dt} = nP_i \left(1 - \frac{P_i}{P_{max}}\right) - k_s P_i X_i \tag{2.1}
$$

$$
\frac{dX_i}{dt} = sX_i \frac{P_i}{\phi_s + P_i}
$$

$$
\frac{dZ}{dt} = \frac{Z - \mu}{\phi_z + P}
$$

where $r$ is the pathogen growth rate, $P_{max}$ is the pathogen carrying capacity, $k_s(k_s)$ is the rate of clearance of pathogen by the variant-specific (cross-reactive) immune response, $s$ is the maximum growth rate of the immune response, $\phi_s(\phi_z)$ is the amount of pathogen that stimulates half-maximal growth of the variant-specific (cross-reactive) immune response, $\mu$ is the rate of stochastic switching among variants and $P = \sum P_i$ is the total density of the pathogen. Pathogens can switch variants at any time, unrelated to growth. We fix the initial conditions to be $P_i(0) = X_i(0) = Z(0) = 1$ and rescale parameters as necessary. As shown previously [33], biologically reasonable parameter ranges satisfy:

$$
k_s < 1 \quad \text{and} \quad r, s \ll \phi_s \leq P_{max}.
$$

(b) Effect of immunogenicity on control
We begin by considering how the time required to control an infection depends on the immunogenicity of the relevant antigens, which, in turn, affects the dynamics of the corresponding immune responses. Immunogenicity is not always well defined, in part because the factors contributing to immunogenicity are not completely understood. Key properties that affect immunogenicity include: antigen density, affinity, the time required for immunogenicity to become manifest, which affects the ease of

Figure 1. A schematic of antigenic variation. We consider both variable dominant antigens that elicit specific immune responses (coloured) and invariant subdominant antigens that elicit a cross-reactive immune response (black). As the variant-specific immune response clears pathogen displaying the dominant red-square antigen, a new variant arises that evades this immune response by displaying a different-dominant antigen (blue triangles). This pattern repeats.

The invariant antigen (black shapes) remains constant across both pathogen variants and we might expect the cross-reactive immune response (black dashed line) to increase over time—albeit slower than the response to the dominant antigen. The question of interest is how rapidly the cross-reactive immune response is able to control the pathogen.
stimulation; \( n_3 \), which affects killing rate and \( n_0 \), which scales the number of naive precursor cells for the cross-reactive response relative to the variant-specific responses. Biologically, these variables are likely correlated. For instance, a change in antigen density by a factor of \( n \) likely would lead to \( n_3 = n_0 = n \) since both stimulation and killing involve a similar process of binding between an immune receptor and an epitope. We can use these three parameters to define the \( k \) and \( \phi \) parameters for cross-reactive responses from equation (2.1) in terms of the \( k_3 \) and \( \phi_3 \) parameters for variant-specific immune response:

\[
\begin{align*}
\phi_3 &= n_0 \phi_3 \\
k_3 &= \frac{k_3}{n_k n_0}
\end{align*}
\]

(2.2)

Note that changing the precursor frequency is algebraically equivalent to changing the killing rate.

We now derive an analytical approximation that allows us to examine how the time required to control the pathogen depends on these three components of immunogenicity. Consider a model that contains only a single antigen and its corresponding immune response:

\[
\begin{align*}
\frac{dP}{dt} &= rP - kPX \\
\frac{dX}{dt} &= sX \left( \frac{P}{\phi + P} \right).
\end{align*}
\]

The immune response will control and clear the pathogen when \( dP/dt \leq 0 \), which occurs when \( X \geq r/k \). The time required to generate this magnitude of a response can be split into two consecutive periods: \( t_1 \), the time until the immune response begins to see significant stimulation (when \( P > \phi \)), and \( t_2 \), the time for the immune response to grow big enough to control the infection (\( X = r/k \)). As shown previously [33], \( t_1 \approx \log(\phi)/r \) and \( t_2 \approx \log(r/k)/s \).

Given this analytic result, we want to see how much the time to control changes between a more immunodominant variant antigen, \( x_i \), and a subdominant invariant antigen, \( z \). If the immune system only responds to the immunodominant antigen, then the time to control will be \( t_i \approx t_1 + t_2 \), if \( k = k_3 \) and \( \phi = \phi_3 \). However, if the immune system only responds to the subdominant invariant antigen, then we see the time to control will be longer (\( t_z > t_3 \)) by a factor of approximately:

\[
\frac{t_z}{t_i} \approx 1 + \frac{(1/r) \log n_0}{t_2} + \frac{(1/s) \log(n_3 n_0)}{t_2}.
\]

(2.3)

As can be seen from figure 2a, this equation implies that even if an invariant subdominant antigen is an order of magnitude less immunogenic than a varying dominant antigen, the antigenic variation will only gain the pathogen a little more time before clearance (specifically, \( O(\log(\max(n_0 n_3) / n_0 n_3)) \)). We confirm this analytic approximation by simulating under the basic model (equation (2.1)) and comparing the time to control with low immunodominance versus the time to control with high immunodominance (figure 2b versus c). As expected from the analytic result, even greatly increased immunodominance leads to little change in the duration of infection.

Despite these model-based results suggesting that antigenic variation should not lead to extended infections, empirical data show that antigenic variation can successfully prolong the duration of infections. This effect can be clearly seen in trypanosomiasis and malaria, both of which can be maintained at high levels for many months [1,3]. Thus, the basic model presented in equation (2.1) must lack a critical component or make an unwarranted approximation. We explore two modifications to this model in the subsequent sections.

(c) Modification of killing term

One significant approximation in the basic model is that immune responses kill pathogens by mass action. However, if antigen is present at very low densities—as might be the case with subdominant invariant antigens—then the rate of killing should saturate.

We now change the classic mass action term for killing of pathogen \( P \) by immune response \( X \) from \( kPX \) to a term that includes saturation: \( cPX/(\theta + X) \). In order for these two terms to converge in the absence of saturation, we define \( c = k\theta \).

With the inclusion of saturation in killing, \( dP/dt \) in equation (2.1) becomes:

\[
\begin{align*}
\frac{dP}{dt} &= rP \left( 1 - \frac{P}{P_{\text{max}}} \right) - c \left( \frac{P}{\theta + X} \right) \\
&= c \left( \frac{P}{\theta + X} \right) - c_P \left( \frac{P}{\theta + X} \right).
\end{align*}
\]

(2.4)

As before, the rate constants for killing depend on the antigen immunogenicity, with the response to subdominant antigens having relatively lower rates of killing and earlier saturation as \( n_k \) increases:

\[
\begin{align*}
\theta_z &= \frac{\theta_x}{n_k} \\
c_z &= k_z \theta_z = \frac{k_z \theta_x}{n_k} = \frac{k_z n_3}{n_k}
\end{align*}
\]

(2.5)

Thus, the ability of the cross-reactive response to kill the pathogen is reduced by a factor of \( n_k^2 \) relative to the variant-specific response. As the immunogenicity of the invariant antigen decreases, a threshold will be crossed when \( c_z < r \) and the cross-reactive response alone will not be able to control the pathogen.

The effect of this threshold can be seen in the simulations shown in figure 3. Figure 3a shows how the cross-reactive response clears the infection when \( c_z > r \), and Figure 3b shows the ineffectiveness of the cross-reactive response at controlling the infection when the subdominant antigen is at sufficiently low levels such that \( c_z < r \).

(d) Incorporation of immune exhaustion

In addition to the assumption of mass action killing, the basic model also assumes the immune response will continue growing as long as antigen persists. However, recent experimental work on immune exhaustion suggests that this standard term for the expansion of immune cells may not be appropriate. We extend our model to incorporate immune exhaustion as previously described in the context of modelling lymphocytic choriomeningitis.
briefly, we add new variables to track the level
of exhaustion for each population of antigen-specific,
and cross-reactive immune cells, \( R \). Exhaustion leads to a
decay in the magnitude of the immune response in a flex-
able manner captured by a Hill function with exponent \( h \)
and coefficient \( q \), such that the maximum decay rate is \( \delta \).
Exhaustion itself increases according to antigen stimulus
(i.e. \( P_i/(\phi_i + P_i) \)) and decays at rate \( \delta_P \). With these
additions, the revised equations describing the dynamics
of the pathogen and immune responses are:

\[
\begin{align*}
\frac{dP}{dt} &= rP \left( 1 - \frac{P}{P_{\text{max}}} \right) - kPXi,
\frac{dXi}{dt} &= sXi + \frac{P_i}{\phi_i + P_i} - \deltaXi \frac{Q^i}{\phi_i^2 + Q^i},
\frac{dZ}{dt} &= sZ \frac{P}{\phi_i + P} - \deltaZ \frac{R^i}{\phi_i^2 + R^i},
\frac{dQ_i}{dt} &= \frac{P_i}{\phi_i + P} - \delta_i Q_i,
\text{and} \quad \frac{dR}{dt} &= \frac{P}{\phi_i + P} - \delta_i R.
\end{align*}
\]

This revised model allows us to explore the synergistic
effects of immunodominance and exhaustion during
antigenically varying infections. Similar to the above
results with saturation in killing, we find two distinct
parameter regimes (figure 4). When the variable antigens
exhibit a low level of immunodominance, the cross-
reactive immune response grows rapidly and controls
the infection (figure 4). Crucially, infection is cleared
prior to exhaustion of any of the responses. In contrast,
when the variable antigens exhibit a high level of
immunodominance, we see the duration of the infection
being extended sufficiently such that the cross-reactive
immune responses become exhausted and the pathogen
persists for an extended period of time (figure 4). The detailed
dynamics of the different variants along
with the variant-specific and cross-reactive immune
responses are shown in figure 5. Note that only the
cross-reactive responses become exhausted—the immunodominant
responses are not exhausted as they are
only transiently stimulated before they rapidly clear the
relevant variant.

These results suggest a prediction: strategic treatment of
an infection by an antigenically varying pathogen—prior to

---

Figure 2. The puzzle: cross-reactive responses should be able to control the infection. Our analytic approximation in equation
(2.3) shows how the time to control depends on antigen immunogenicity. In (a), we plot this function while varying immunogenicity owing to antigen abundance (which equally affects stimulation \( n_g \) and killing \( n_h \) with \( \phi = n_g10^3 \) and \( k = 10^{-5}/n_h \)) and the number of naive precursor cells \( n_0 \). We see that the relative time to control changes only modestly (less than a factor of 2)
even when the subdominant antigen is 1000-fold less abundant than the dominant antigen. In (b,c), we plot simulated trajectories from our basic model in which the invariant antigen either exhibits low immunodominance ((b); \( n_g = n_h = 2 \)) or high immunodominance ((c); \( n_g = n_h = 1000 \)). As predicted by the analytic result in (a), even greatly increasing the immunodominance leads to relatively little increase in the duration of infection. Parameters: \( r = 2, P_{\text{max}} = 10^7, k = 10^{-5}, \phi_c = 10^3, n_0 = 1, k_z = k/n_c, \phi_z = n_c/\phi_z, s = 1.3, \mu = 10^{-8} \).

---

virus [14]. Briefly, we add new variables to track the level
of exhaustion for each population of antigen-specific, \( Q_i \)
and cross-reactive immune cells, \( R \). Exhaustion leads to a
decay in the magnitude of the immune response in a flex-
able manner captured by a Hill function with exponent \( h \)
and coefficient \( q \), such that the maximum decay rate is \( \delta \).
Exhaustion itself increases according to antigen stimulus
(i.e. \( P_i/(\phi_i + P_i) \)) and decays at rate \( \delta_P \). With these
additions, the revised equations describing the dynamics
of the pathogen and immune responses are:
exhaustion of the cross-reactive responses—could lead to long-lasting cross-reactive immunity.

(c) Dynamics of treatment

Given our prediction above, we now explore the consequences of antimicrobial drug treatment of infections under both of our hypotheses—the first postulating saturation in immune killing and the second including exhaustion of immune responses. We model antimicrobial drug treatment by including a mass action term for killing of the parasite by the drug. The treatment begins at time $t_D$, and a constant drug concentration $D$ is maintained until the pathogen is cleared. The inclusion of treatment in both models results in rapid clearance of the infection. We find, however, that treatment has very different consequences depending on the model for the generation of cross-reactive immunity and thus the dynamics of secondary infections.

We first consider the hypothesis of saturation in killing. From the left panels in figure 6, we see that antimicrobial treatment has little long-term benefit under this model, regardless of when treatment is applied. Cross-reactive immunity following the primary infection grows higher if treatment is delayed, but this immunity is fundamentally unable to control a secondary infection owing to saturation in its ability to kill.

In contrast, under the exhaustion hypothesis, we find that the outcome depends strongly on the timing of treatment (right panels in figure 6). Maximal cross-reactive immunity arises when treatment occurs at intermediate times after infection (middle-right panel), when the cross-reactive immunity reaches a level sufficient to confer protection and rapidly clear a secondary infection. However, both early and late treatment (top and bottom rows) yield limited cross-reactive immunity and little, if any, protection from secondary infection.

Intuitively, early treatment gives too little exposure to invariant antigens to generate significant cross-reactive immunity, and late treatment results in little immunity because exhaustion has already set in prior to treatment. Maximal generation of cross-reactive immunity occurs at a time that balances these two extremes by providing sufficient antigen exposure to generate significant immunity but not so much as to trigger exhaustion.
We now focus our attention on the few papers that have explicitly considered the effect of cross-reactive immune responses on the dynamics of persistent infections [19,24]. The results of Antia et al. [19] are based on the assumption that immune memory requires persistent antigen and hence analysis steady-state behaviour arising from predator–prey dynamics of immune responses with a significant decay term. Since the time of this publication, it has become clear that this approximation does not hold as immunological memory can be long-lived in the absence of antigen [42]. The model by Recker et al. [24] considered the interplay between variant-specific and cross-reactive immune responses. They constructed a model for malaria infection in which variant-specific immune responses were long-lasting but cross-reactive immune responses were short-lived. This qualitatively distinct behaviour of variant-specific and cross-reactive responses was necessary in order to obtain orchestrated appearance of variants; however, the model did not supply an underlying reason for why the cross-reactive immune responses might be short-lived. We note that our model with immune exhaustion provides an explanation for this disparity in longevity between the variant-specific and cross-reactive immune responses in the model by Recker. Our model goes further than simply providing a mechanistic explanation. As described in the section on 'Dynamics of treatment', our immune exhaustion hypothesis predicts that treatment can lead to the generation of protective, long-lasting cross-reactive immunity. In contrast, the Recker hypothesis predicts that treatment should not extend the longevity of cross-reactive responses.

We know of at least two persistent infections with experimental evidence consistent with our hypotheses. For trypanosome infections, a vaccine derived from an antigen associated with the non-varying *Trypanosoma brucei rhodesiense* flagella has been found to provide partial...
heterologous protection in cattle [43]. However, the flagella antigens are present at very low copy number with the dominant variable antigens (the variable surface glycoproteins or VSGs) covering almost the entire cell. Thus, it is not clear whether natural cross-reactive antibodies to trypanosomes fail to clear the infection because immune exhaustion prevents the production of sufficiently high concentrations of antibodies to the flagella or because low density of flagella antigens leads to saturation in killing. For malaria infections, cross-reactive antibodies have occasionally been observed in natural human infections of \textit{Plasmodium falciparum} [44]. Intriguingly, the data tentatively suggest that these antibodies might not occur in children in malaria-endemic regions [44]—which could be consistent with immune exhaustion of these responses at a very young age. In addition, an increasing number of experimental studies have demonstrated the potential to confer cellular-based immunity against malaria by infecting people or mice with low doses of live or killed parasites followed by early treatment [45–47]. The specific mechanism of action for these successes remains unknown, although exhaustion has been mentioned as one possibility for the lack of natural immunity [48]. Finally, a very recent paper has provided direct evidence that CD4 cells in humans infected with malaria show signs of exhaustion and that, at least in


graphical content

Figure 6. Treatment of a primary infection can result in protective immunity against secondary infections. Our two hypotheses (saturation in killing and immune exhaustion) make different predictions. We show how the time of treatment of primary infection affects the level of cross-reactive immunity and the consequences for a secondary infection at day 120. Under the saturation hypothesis (left panels), treatment does not lead to the generation of protective immunity, regardless of the time of treatment. Under the exhaustion hypothesis (right panels), appropriately timed treatment can lead to the generation of protective cross-reactive immunity (middle-right panel). In this scenario, treatment too early clears the primary infection before sufficient cross-reactive immunity can be generated, and treatment too late results in exhaustion of cross-reactive immunity. Parameters: $n_k = n_b = 10^3$, $D = 20$. Other parameters same as figure 4.
mice, disrupting the exhaustion pathway speeds clearance of the infection [15].

While our hypotheses explaining antigenic variation are intriguing, our study has limitations. For instance, our explanations will not apply to diseases such as HIV for which variation may be a consequence rather than a cause of persistence. Further, disease dynamics may be influenced by factors such as competition between epitopes for limited antigen presentation resources. Finally, none of these hypotheses are mutually exclusive, so the true answer may be some combination of explanations.

Regardless of limitations, exploratory models are still useful if they facilitate the generation of testable predictions that can discriminate between different hypotheses. In our case, the key predictions involve the effect of treatment. Only the immune exhaustion hypothesis predicts treatment can lead to the generation of protective, long-lasting cross-reactive immunity. Both the Recker hypothesis and our saturation hypothesis predict that treatment should not substantially affect the dynamics of subsequent infections.

We end by noting that if immune exhaustion plays a key role in the dynamics of persistent infections, then appropriately timed treatment of primary infections could be used as a form of vaccination.

We thank Vitaly Ganusov for helpful discussion. This research was supported in part by NIH grant R01 AI049334 to R. Antia and NIH grants R01 AI030048 and P01 AI080192 to R. Ahmed. B.F.K. was supported by the Fannie and John Hertz Foundation.

REFERENCES


Antigenic variation versus cross-immunity  P. L. F. Johnson et al. 2785


