

# How a well-adapted immune system is organized

Andreas Mayer<sup>a</sup>, Vijay Balasubramanian<sup>b,c</sup>, Thierry Mora<sup>d,1</sup>, and Aleksandra M. Walczak<sup>a</sup>

<sup>a</sup>Laboratoire de Physique Théorique and <sup>d</sup>Laboratoire de Physique Statistique, CNRS, Université Pierre et Marie Curie and École Normale Supérieure, 75005 Paris, France; <sup>b</sup>Department of Physics and Astronomy, University of Pennsylvania, Philadelphia, PA 19104; and <sup>c</sup>Initiative for the Theoretical Sciences, The Graduate Center, The City University of New York, New York, NY 10016

Edited by José N. Onuchic, Rice University, Houston, TX, and approved March 3, 2015 (received for review November 14, 2014)

**The repertoire of lymphocyte receptors in the adaptive immune system protects organisms from diverse pathogens. A well-adapted repertoire should be tuned to the pathogenic environment to reduce the cost of infections. We develop a general framework for predicting the optimal repertoire that minimizes the cost of infections contracted from a given distribution of pathogens. The theory predicts that the immune system will have more receptors for rare antigens than expected from the frequency of encounters; individuals exposed to the same infections will have sparse repertoires that are largely different, but nevertheless exploit cross-reactivity to provide the same coverage of antigens; and the optimal repertoires can be reached via the dynamics of competitive binding of antigens by receptors and selective amplification of stimulated receptors. Our results follow from a tension between the statistics of pathogen detection, which favor a broader receptor distribution, and the effects of cross-reactivity, which tend to concentrate the optimal repertoire onto a few highly abundant clones. Our predictions can be tested in high-throughput surveys of receptor and pathogen diversity.**

immune repertoires | optimal coding | repertoire diversity | competitive exclusion | jamming

The adaptive immune system protects organisms from a great variety of pathogens by maintaining a population of specialized cells, each specific to particular challenges. Together these cells cover the array of potential threats. To recognize pathogens, the immune system relies on receptor proteins expressed on the surface of its main constituents, the B and T lymphocytes. These receptors interact with antigens (small molecular elements making up pathogens), recognize them through specific binding, and initiate the immune response. Each lymphocyte expresses a unique receptor formed from random combinations encoded in the genome. The receptors later undergo selection through the death and division of the lymphocytes that express them, as well as mutations in the case of B lymphocytes. The diversity of the receptor repertoire determines the range of threats that the adaptive immune system can target.

The detailed composition of the immune receptor repertoire, and not just its breadth, is important for conferring effective protection against infections. Broadly speaking, a diverse population of receptors will confer wider immunity, and a larger clonal population of a particular receptor will confer more effective immunity against the pathogens to which it is specific. However, there is a trade-off between diversity and clone sizes because the number of receptors is limited. By selectively proliferating some receptors at the expense of others, the immune system retains a memory of past infections (1), facilitating subsequent immune responses. Furthermore, although infections increase the populations of receptors with the greatest specificity, they can also lead to a reorganization of the immune repertoire as a whole (2).

How should the repertoire be organized to minimize the cost of infections? We develop a framework for answering this question by abstracting key general features of the adaptive immune system: The receptor repertoire is bounded in size, receptors are “cross-reactive” (each antigen binds many receptors; each receptor binds many antigens), and the cost of an infection increases with time. Given these general assumptions, we consider a simplified landscape of pathogens, where infections are

drawn from a fixed distribution. By simplifying the setting in this way, and independently of the detailed dynamics of immune responses, we arrive at broad insights about the composition of immune repertoires that are optimal for their pathogenic environments. Our framework is not meant to give a complete account of immunity. To do so we would need to include several other components of the immune system, such as interaction between its innate and adaptive arms and avoidance of autoimmunity. The latter problem—the challenge of discriminating self from nonself—has been the focus of many theoretical studies of the immune system (3, 4). This paper primarily investigates the relation between the adaptive repertoire and the pathogenic environment, but we also discuss how other components and constraints of the immune system can be incorporated into our model.

The theory predicts, counterintuitively, that the number of receptors specific to rare pathogens will be amplified relative to the probability of encounter, at the expense of receptors for common infections. We also find that two organisms responding to a pathogen distribution will display unique populations of immune receptors, even though their coverage of pathogens will be similar. How can the immune system achieve these sorts of optima? Surprisingly, we find that simple competition between receptor clones can drive the population to the optimal composition for minimizing the cost of infections.

New high-throughput methods are making it possible to survey B-cell and T-cell receptor diversity in fish (5, 6), in mice (2, 7), and in humans (8–11). As methods are developed to better characterize pathogenic landscapes and receptor cross-reactivity, predictions for the composition of optimal repertoires derived from our framework can be directly compared with experiments. To arrive at our results we ask how the immune system should be organized to perform its function well, rather than starting with the detailed dynamics of its components. We are proposing that the universal features of the adaptive immune system follow simply from general statistical considerations, whereas the detailed

## Significance

**The adaptive immune system uses the experience of past infections to prepare its limited repertoire of specialized receptors to protect organisms from future threats. What is the best way of doing this? Building a theoretical framework from first principles, we predict the composition of receptor repertoires that are optimally adapted to minimize the cost of infections from a given pathogenic environment. A naive repertoire can reach these optima through a biologically plausible competitive mechanism. Our findings explain how limited populations of immune receptors can self-organize to provide effective immunity against highly diverse pathogens. Our results also inform the design and interpretation of experiments surveying immune repertoires.**

Author contributions: A.M., V.B., T.M., and A.M.W. designed research, performed research, contributed new reagents/analytic tools, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

<sup>1</sup>To whom correspondence should be addressed. Email: thierry.mora@gmail.com.

This article contains supporting information online at [www.pnas.org/lookup/suppl/doi:10.1073/pnas.1421827112/-DCSupplemental](http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1421827112/-DCSupplemental).

dynamical implementation arises from the historical contingencies of evolution.

### Definition of the Problem

To find the optimal repertoire distribution we must consider the nature of antigen–receptor interactions and a penalty that the immune system pays for not recognizing antigens. This penalty must reflect the facts that recognition should happen within a reasonable time, before the pathogen colony can significantly increase its size; the interactions between the immune receptors and antigen are probabilistic; and not all antigens are equally frequent. We assume that, although the immune system cannot predict precisely which antigens it will encounter and when, it incorporates an estimate of the probabilities of their occurrences. We also take these probabilities to be constant in time. This is an idealization grounded in a separation of timescales, which assumes the distribution of antigens remains constant on timescales on which the immune system adapts.

The above ideas are the basis for our cost function, which reflects the penalty of nonrecognition, for a given repertoire and antigenic environment. In Fig. 1 we introduce a quantification of the problem. Given  $Q_a$ , the probability that the next infection will be caused by antigen  $a$ , we model the immune repertoire by a distribution of receptors  $P_r$ , from which lymphocytes with the receptor  $r$  are drawn at random.

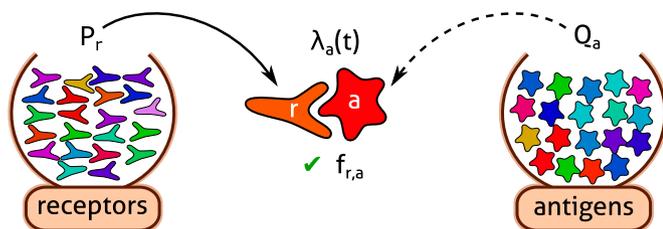
An antigen  $a$  and a receptor  $r$  interact with a certain strength set by the binding affinity between the two molecules. This is described by the probability  $f_{r,a}$  (which we call the “cross-reactivity function”) that an antigen  $a$  encountering a receptor  $r$  results in a recognition event, leading to the activation of the lymphocyte expressing that receptor. Each encounter of the antigen  $a$  with a random receptor has a probability  $\bar{P}_a = \sum_r f_{r,a} P_r$  to lead to recognition and trigger an immune response. Thus,  $\bar{P}_a$  can be viewed as the coverage of antigen  $a$  by the repertoire.

Given this coverage, we consider  $\bar{F}_a$ —the average harm caused by antigen  $a$ . We show below that, consistent with intuition,  $\bar{F}_a$  is a decreasing function of the coverage  $\bar{P}_a$ . The overall expected cost is then just the harm averaged over the antigen distribution:

$$\text{Cost}(\{P_r\}) = \langle F \rangle = \sum_a Q_a \bar{F}_a(\bar{P}_a). \quad [1]$$

The need to defend against many antigens at the same time with a limited number of receptors introduces a trade-off. If more receptors recognize an antigen, there are less to protect against other threats.

Finally, we derive an expression for the average harm  $\bar{F}_a$  caused by antigen  $a$ . During its time in the periphery, an antigen  $a$  will encounter and possibly interact with receptors at a rate  $\lambda_a(t)$  that increases with time as the pathogen population grows. Each encounter will occur with a different receptor  $r$  drawn from  $P_r$ . The mean number of encounters between antigens and receptors after a time  $t$ , which we call effective time, is defined as  $m_a(t) = \int_0^t d\tau \lambda_a(\tau)$ , where  $t=0$  is set by the introduction of the



**Fig. 1.** Schematic of a statistical model of antigen recognition by the adaptive immune system. After infection, antigen  $a$  encounters immune receptor  $r$  at random with a rate  $\lambda_a(t)$ . An encounter leads to a successful recognition with a probability  $f_{r,a}$  that reflects the matching between a given antigen–receptor pair.

antigen. The time  $t$  to the first recognition event, or response time, is random and depends on the coverage  $\bar{P}_a$ .

The longer the system fails to detect the antigen, the more likely the infection is to become harmful. We assume that the integrated harm caused by an antigen since the beginning of an infection is an increasing function  $F_a(t)$  of the time of first recognition. How exactly  $F_a$  grows with time may strongly depend on the type of infection and receptors (12–14). The mean harm inflicted to the organism by the attack of an antigen  $a$  is then given by this quantity averaged over the distribution of possible response times,  $\bar{F}_a = \bar{P}_a \int_0^{+\infty} dm F_a[t_a(m)] e^{-m\bar{P}_a}$ , where  $t_a(m)$ , the inverse function of  $m_a(t)$ , is the amount of time it takes for  $m$  encounters to occur between the receptors and pathogen  $a$  (see *SI Text, Appendix A* for a derivation). The result depends on the cost expressed as a function of the effective time  $m$ ,  $F_a[t_a(m)]$ , which we denote  $F_a(m)$  to simplify notation. We consider several specific choices of this effective cost function in *Results*.

Our aim here is to propose a general framework for thinking about the repertoire. Thus, we do not explicitly model intracellular communication, cell differentiation, activation of cofactors, coordination of different cell types, the interaction with the innate immune system, and the full complexity of the recognition process. The idea is that  $F_a(m)$  implicitly summarizes all of these factors in terms of an effective cost.

In general the cost function  $F_a(m)$  depends on the antigen  $a$ , reflecting the various virulences of different pathogens. To simplify, we can assume that the cost function takes the factorized form  $F_a(m) = \mu_a F(m)$ , where  $\mu_a$  is the pathogen-dependent virulence factor, and  $F(m)$  describes how all threats develop with time. The cost will then take the form  $\sum_a \mu_a Q_a \bar{P}_a \int_0^{+\infty} dm F(m) e^{-m\bar{P}_a}$ . In this expression, the virulence factor  $\mu_a$  of a pathogen plays the same role as its likelihood  $Q_a$ . Some pathogens are rare but very virulent (like anthrax), whereas others may be common but not very virulent (like the common cold), and an ideal immune system should be able to cope with both. In our model the overall “threat” of a pathogen is expressed as the product of the two,  $\mu_a Q_a$ . In practice  $\mu_a$  can be absorbed into the definition of  $Q_a$  and is omitted in the rest of this paper.

Given such a model of the recognition process, there exists an optimal adaptive immune system, characterized by the choice of the receptor distribution  $P_r$ , that minimizes the expected cost in a given antigenic environment  $Q_a$ . The optimal repertoire is found by minimizing the expected cost in Eq. 1 with respect to  $P_r$ , subject to constraints of nonnegativity ( $P_r \geq 0$ ) and normalization ( $\sum_r P_r = 1$ ). Simple local extremality conditions are sufficient for optimality because our problem can be shown to be convex (*SI Text, Appendix B*). The condition  $\sum_r P_r = 1$  is a normalized version of the constraint that the total number of receptors is limited.

### Results

#### The Optimal Repertoire Is More Uniform Than the Pathogen Distribution.

We can now ask how best to distribute the receptors to minimize the cost (Eq. 1) for a given antigenic environment. To begin, we neglect cross-reactivity (later we will see that this is equivalent to looking at the structure of the repertoire at scales larger than the cross-reactivity). In this case antigens and receptors can be associated one by one by a cross-reactivity function  $f_{r,a} = 1$  if  $r = a$  and 0 otherwise. In this case we can analytically determine the optimal distribution (*SI Text, Appendix D, section 2*)  $P_r^* = \max[\bar{F}'^{(-1)}(-\lambda/Q_r), 0]$ , where  $\bar{F}'^{(-1)}$  denotes the inverse function of the derivative of  $\bar{F}_a = \bar{F}(\bar{P}_a)$  expressed as a function of  $\bar{P}_a$ , and  $\lambda$  is a positive constant fixed by the normalization  $\sum_r P_r^* = 1$ . Table 1 presents results for several representative cost functions.

A simple scenario occurs when the pathogen population grows exponentially in time, as do the cost and the encounter rate—reflecting the proliferative nature of pathogens. In this case the cost grows linearly in the number of encounters; i.e.,  $F(m) = m$  (*SI Text, Appendix C*). Then we find that the optimal fraction of the repertoire taken up by a given receptor is proportional to the

square root of the threat (combination of frequency and virulence) of the corresponding antigen  $P_r^* \propto \sqrt{Q_r}$ . Intuitively, we expect that the optimal repertoire should focus its resources on receptors recognizing the most common or virulent antigens. However, this enhanced protection against frequent or virulent antigens comes at the cost of a slower response against the uncommon and harmless antigens, and this bias toward more threatening antigens must remain limited. The square-root dependence reflects a particular trade-off between these two opposing constraints, by directing more resources toward more threatening antigens while uniformizing the distribution compared with a linear dependence. Intriguingly, the same square-root dependence has been found as an optimal solution for the size of tRNA pools as a function of codon use (15) and in a model for the screening of suspicious individuals (16).

The extent to which more resources are directed toward more threatening antigens depends on the relative gains and losses of earlier and later recognition events, which are captured in our model by the effective cost function  $F(m)$ . In general, steeper cost functions imply more flattened distributions of receptors. The cost function  $F(m) = m^\alpha$ , and its associated optimal distribution  $P_r^* \propto Q_r^{1/(1+\alpha)}$ , help illustrate this point. Such cost functions can arise when both  $m(t)$  and  $F(t)$  increase exponentially as a function of time, but with different exponents (SI Text, Appendix C). When  $\alpha$  is large, the cost of nonrecognition increases very quickly with time, calling for an urgent response. Consequently the optimal immune system tends to cover the space uniformly to get all potential threats, even the unlikely ones, under control. Conversely, when  $\alpha$  is small, the harm caused by pathogens does not explode with time, permitting the system to recognize the rarer pathogens late and focus its resources on the common ones.

In some situations, there may be little or even no difference between a late response and no response at all, because the total harm caused by an infection stabilizes. For example, consider the cost  $F(m) = 1 - e^{-\beta m}$  that saturates at large effective times. In this case, the optimal solution (Table 1) relates receptor and antigen through a square root as for linear cost, but with a cutoff at low  $Q_a$ . This cutoff occurs because there is little benefit to having receptors recognizing rare or harmless antigens, whose recognition is likely to happen late, when differences in recognition times do not matter anymore. This result is consistent with the observation that the immune system may sometimes ignore infections if the harm they cause is too small, as in, e.g., the case of simian immunodeficiency virus infection of sooty mangabey (17).

Real harm may occur only when the effective time  $m$  crosses a threshold. This situation can be modeled by taking  $F(m) = \Theta(m - m_0) = 0$  for  $m < m_0$  and 1 otherwise. In this case the receptor distribution is organized to maximize the chance of detection before  $m_0$ . The optimal repertoire for this cost (Table 1) has  $n$  receptors for the least threatening pathogens (cutoff at

low probabilities) and a drastically flattened receptor distribution (logarithm of the pathogen distribution).

Is there a cost function for which the receptor distribution is not flattened relative to the pathogen distribution? This occurs in a special case where cost increases very slowly (logarithmically) with effective time. However, in general, cost is minimized by a receptor distribution that is flattened relative to the pathogen distribution.

**Cross-Reactivity Dramatically Reduces Diversity in the Optimal Repertoire.** By allowing receptors to bind to a variety of antigens, cross-reactivity should permit the immune system to reduce the number of receptor types required to cover the whole range of possible threats. We show that given sufficient cross-reactivity, the optimal immune repertoire concentrates all its resources on a few receptors, which together tile antigenic space.

Following Perelson and Oster (4), we think of receptors and antigens as points in a common high-dimensional shape space, whose coordinates are associated to unspecified physicochemical properties. For simplicity, assume that cross-reactivity depends only on the relative position of receptor and antigen in shape space  $f_{r,a} = f(r - a)$ , where  $f$  is a decreasing function of the distance between  $a$  and  $r$ . Short distances in shape space correspond to a good fit between the two molecules, leading to strong recognition, whereas long distances translate into weak interactions and poor recognition.

To build intuition, we first consider an analytically solvable example (Fig. 2). We describe the space of receptors and antigens by a single continuous number and assume a Gaussian antigen distribution with variance  $\sigma_a^2$  and Gaussian cross-reactivity of width  $\sigma$ , which sets the typical distance within which a receptor and antigen interact. We derive the optimal receptor distributions for costs of the form  $F(m) = m^\alpha$  (SI Text, Appendix D, section 3b). For narrow cross-reactivities ( $\sigma < \sigma_c = \sigma_a \sqrt{1 + \alpha}$ ), the optimal receptor distribution is Gaussian with variance  $(1 + \alpha)\sigma_a^2 - \sigma^2$  and the optimal cost is independent of  $\sigma$ . For wide cross-reactivities ( $\sigma > \sigma_c$ ), the receptors are optimally of a single type with reactivity centered on the pathogen distribution, whereas the optimal normalized cost increases with  $\sigma$  because the receptor is unnecessarily broadly reactive. These results arise from a tension between two opposing tendencies. As in the non-cross-reactive case, the need to cover rare pathogens broadens the optimal receptor distribution relative to the pathogen distribution. However, cross-reactivity has the opposite effect, favoring more concentrated distributions.

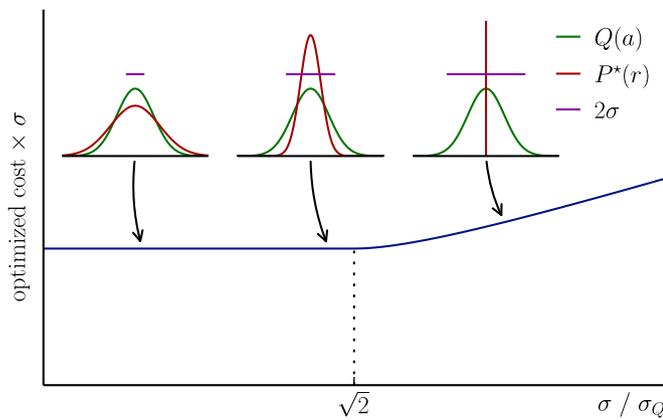
Does cross-reactivity generically drive the optimal receptor distribution to cluster into peaks? We investigated this question numerically. For concreteness, we consider a linear cost  $F(m) = m$  and random pathogen environments in one or two dimensions constructed by drawing each  $Q_a$  from a log-normal distribution characterized by a coefficient of variation  $\kappa$ . The shape space is taken to be bounded and discretized, and we use accelerated gradient projection optimization (SI Text, Appendix E). We find that the optimal repertoire  $P^*$  is strongly peaked on a discrete forest of receptors (Fig. 3 A and B). The width of these peaks decreases as numerical precision is increased, suggesting that in a continuous limit the optimum consists of a weighted sum of Dirac delta functions, i.e., distinct, discretely spaced receptors in different amounts (Fig. S1).

By inspection, the peaks tend to repel each other and to organize into local tiling patterns, as further evidenced by the damped oscillations in the radial distribution function (18) (SI Text, Appendix F and Fig. S2A). This exclusion is a sensible way to distribute resources, as it limits redundant protection against the same pathogens. The spacing between peaks roughly follows the cross-reactivity scale  $\sigma$ , suggesting that  $P^*$  is smooth when viewed at scales larger than  $\sigma$ . Confirming this,  $\tilde{P}^*$  (i.e., the coverage of the antigenic space by the receptors) smoothly tracks the variations in the antigen distribution  $Q$  at a broad scale (Fig. 3A). When viewed coarsely in this way, cross-reactivity is irrelevant and  $P^*$  tends to the solutions of Table 1. Indeed, at these broad

**Table 1. Cost functions and optimal repertoires**

$F(m)$	$P_r^*$	$A(\tilde{N}_a)$
$m^\alpha$	$C Q_r^{1/(1+\alpha)}$	$C'(N_{st}/\tilde{N}_a)^{1+\alpha}$
$\ln m$	$C Q_r$	$C'(N_{st}/\tilde{N}_a)$
$1 - \exp(-\beta m)$	$\max\{C\sqrt{Q_r} - \beta, 0\}$	$C' / (\beta + N_a/N_{st})^2$
$\Theta(m - m_0)$	$\max\{\ln(Q_r)/m_0 - C, 0\}$	$C' \exp(-m_0 \tilde{N}_a/N_{st})$

The cost function  $F(m)$  measures the harm caused to an organism by the time that immune receptors have had  $m$  encounters with a pathogen. The optimal receptor distribution  $P^*$  is determined by minimizing this cost, given a pathogen distribution  $Q$  and a cross-reactivity function  $f_{r,a}$  specifying the probability that receptor  $r$  binds to antigen  $a$ . The second column gives the form of  $P^*$  over scales larger than the cross-reactivity. The optimal  $P^*$  can be reached as a steady state resulting from competitive binding between receptors and antigens (last section of Results) quantified by an "availability function"  $A$ .  $\tilde{N}_a = \sum_r N_r f_{r,a}$  represents the coverage of antigen  $a$  by the repertoire;  $N_{st} = \sum_r N_r$  is the total steady-state population; and  $C, C', \beta$ , and  $m_0$  are positive constants.



**Fig. 2.** The optimal cost and receptor distributions for protecting against a one-dimensional Gaussian antigenic landscape  $Q(a)$  of variance  $\sigma_Q^2$ , as a function of the cross-reactivity width  $\sigma$ . As  $\sigma$  increases, the optimal distribution  $P^*(r)$  becomes narrower and narrower (Left and Center Insets), until it concentrates entirely onto a single point, for  $\sigma \geq \sqrt{2}\sigma_Q$  (Right Inset). The minimal cost (multiplied by  $\sigma$  for a comparison at constant recognition capability) is constant below the transition point, but increases with  $\sigma$  past it. The cross-reactivity function, which quantifies the affinity between receptor  $r$  and antigen  $a$  as a function of their distance in shape space, has a Gaussian form,  $f(r-a) = \exp[-(r-a)^2/2\sigma^2]$ , and the cost function is linear in the effective recognition time,  $F(m) = m$ .

scales, the distribution of peaks is uniform, as demonstrated by the very low power in the spectrum of  $P_r$  at small wave vectors (SI Text, Appendix F and Fig. S2B), indicating that the number of receptors contained in any given large area of the shape space is very reproducible. This phenomenon of small-scale randomness with large-scale regularity is called disordered hyperuniformity (19) and arises in jammed packings as evidence of the incompressibility of the material. In biological terms, hyperuniformity means that the distribution of receptor peaks provides a much more uniform coverage of the antigen space than if the peaks were positioned randomly according to a Poisson distribution. For our optimal repertoires small-scale fluctuations get smoothed out by cross-reactivity and can be tolerated, whereas at large scales the fluctuations track variations in the antigenic landscape to provide smooth coverage (Fig. S3).

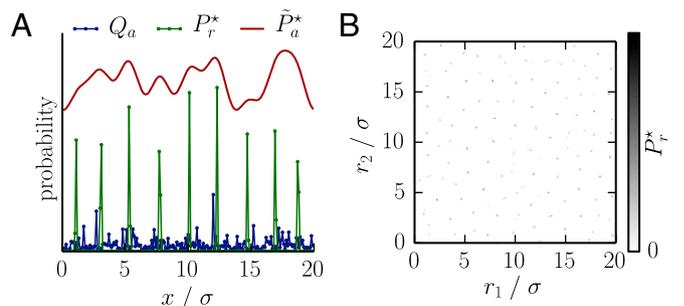
To test the generality of our findings we tested other choices of cross-reactivity functions (SI Text, Appendix G). We have so far assumed a unique scale  $\sigma$  for cross-reactivity, consistent with recent reports that cross-reactivity is local in antigenic space (20). However, receptor-antigen recognition can be very specific and sensitive to single mutations (21, 22) or extremely degenerate across very dissimilar antigens (23). To account for these long-range effects, we also tested fat-tailed cross-reactivity functions. We found that, for a variety of non-Gaussian cross-reactivity functions, long tailed or not, the optimal repertoire remains strongly peaked, although the position, number, and strength of the peaks do change (Fig. S4). Next, to relax the assumption that the cross-reactivity width is uniform across receptors, we tested receptor-dependent cross-reactivities  $\sigma_r$ , drawn from a log-normal distribution. While the regularity of the local tiling structure is affected by this additional heterogeneity (just as we expect in a packing of spheres of variable size), the large-scale hyperuniformity is nonetheless preserved (Fig. S5). Next we considered distributions of antigens with correlations across shape space (reflecting, e.g., phylogenetic correlations between pathogens). Again we find peaked optimal receptor distributions (Fig. S6), similar to those for uncorrelated antigen landscapes. For computational reasons, we restricted our analysis to 2D pathogen landscapes, but the analogy with random packing problems that we discussed above allows us to expect that all of these results will hold generally in higher dimensions. Finally, we incorporated the avoidance of the self by excluding from the

optimization all receptors within distance  $\sigma$  of a set of randomly positioned self-antigens (SI Text, Appendix H and Fig. S7). We find that receptors are likely to be found near the boundary of these exclusion zones, but otherwise keep the same general tiling structure.

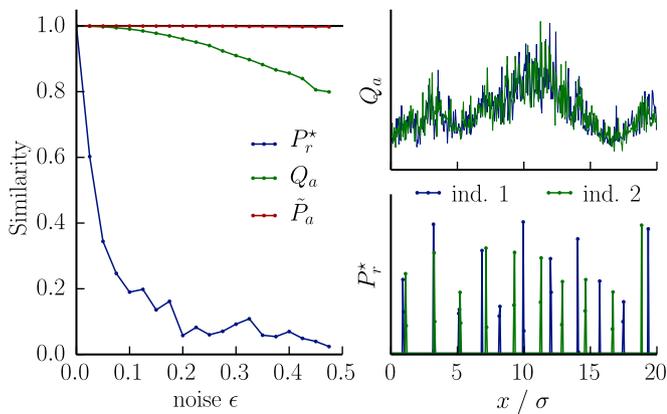
In summary, the optimal immune repertoire looks random at scales smaller than the cross-reactivity, but has the structure of a disordered tiling at larger scales so that, after accounting for cross-reactivity, the repertoire smoothly covers the pathogen landscape. These findings have an important consequence for different individuals exposed to the same pathogenic environment. Each individual will experience a slightly different spectrum of antigens because of the statistics of encounters and other sources of variability. These slightly different experiences of the same world lead to optimal repertoires with a striking property—the receptor distributions are largely different, even though their coverage of the pathogen landscape is similar after including cross-reactivity (Fig. 4). This finding can be compared with surveys of “public” repertoires of immune receptors (2, 24).

**The Optimal Repertoire Can Be Reached Through Competition for Antigens.** The results presented so far have established how repertoires should be structured to provide optimal protection. Given the complex interdependences between receptors arising from local and global trade-offs, one might think that the globally optimal solution could be reached only via some biologically implausible centralized mechanism distributing resources system-wide. In fact, we show that the optimal repertoire can be reached through self-organization, via competitive evolution of receptor populations under antigen stimulation.

We consider a model that is similar to that introduced by de Boer and Perelson (25) and de Boer et al. (26) for competitive dynamics of B and T cells (SI Text, Appendix I). Its main assumptions are that division of receptor-expressing lymphocytes is driven by antigen stimulation and that receptors compete for the limited supply of antigens. At each time step, a random antigen  $a$  is drawn from the distribution  $Q_a$ . Each receptor type  $r$  responds to it by expanding or shrinking its population  $N_r$  according to its specificity, by an amount  $N_r \Delta t [A(\sum_r N_r f_{r,a}) f_{r,a} - d]$ , where  $\Delta t$  is the time step: Receptors proliferate upon successful recognition of antigens (first term) and die with a constant rate  $d$  (second term). In the absence of competition, the proliferation rate should be proportional to  $f_{r,a}$ , but the antigen  $a$  may also bind other receptors, reducing its availability for receptor  $r$ . The coverage of antigen  $a$  by the repertoire,  $\tilde{N}_a = \sum_r N_r f_{r,a}$ , quantifies the breadth of the receptor pool competing to bind with  $a$ . The availability of antigen  $a$  for binding is assumed to be a decreasing function  $A(\tilde{N}_a)$  of its coverage. The stimulation of  $r$  by  $a$  is thus



**Fig. 3.** Cross-reactivity plays an important role in shaping the optimal repertoire, often leading to highly peaked repertoires. (A and B) The optimal receptor distribution  $P_r^*$  for (A) 1D and (B) 2D random environments. Despite being peaked, the optimal distribution of receptors covers the antigenic space fairly uniformly, as shown by its coverage by the receptors,  $\tilde{P}_a^* = \sum_r f_{r,a} P_r^*$ , shown in the one-dimensional case (A). The cross-reactivity and cost functions are the same as in Fig. 2. The antigenic landscape  $Q_a$  is generated randomly from a log-normal distribution with coefficient of variation  $\kappa = 1$ .



**Fig. 4.** Two individuals in the same environment  $Q_a$  that see it with slightly different noises have similar coverages of the antigenic space, but achieve it with different receptors. This results in largely nonoverlapping repertoires. Shown are the overlaps (normalized to be between 0 and 1) between the experienced pathogen distributions  $Q_a$ , the resulting optimal receptor distributions  $P_r^*$ , and the corresponding coverages  $\tilde{P}_a$ , as a function of the noise  $\varepsilon$  with which individuals perceive the environment. *Right* plots show an example of antigenic environments and optimal receptor distributions for  $\varepsilon = 0.2$ . We calculated the optimal receptor distributions for two individuals 1 and 2 experiencing respective environments  $Qe^{z_1}$  and  $Qe^{z_2}$ , where  $Q$  is a random environment with fluctuations on scales larger than the cross-reactivity  $\sigma$  [power spectrum  $\propto 1/(1+(10q\sigma)^2)$ ] normalized so that its coefficient of variation is 0.5, and  $z_1, z_2$  are Gaussian noises of mean zero and variance  $\varepsilon^2$ . The choice of cost and cross-reactivity functions are the same as in Fig. 2.

modified to  $A(\tilde{N}_a)f_{r,a}$  in the equation for the growth rate. In the limit of fast sampling of antigens, or mean-field limit ( $\Delta t \rightarrow 0$ ), these stochastic dynamics are well described by the deterministic differential equations

$$\frac{dN_r}{dt} = N_r \left[ \sum_a Q_a A \left( \sum_{r'} N_{r'} f_{r',a} \right) f_{r,a} - d \right]. \quad [2]$$

For a given pathogenic environment, the total steady-state receptor population size  $N$  will be set by the death rate  $d$ , which counterbalances growth at steady state. Although in reality the ability of the system to reorganize itself diminishes with age, for simplicity we take all rates to be constant.

The stable fixed points of the mean-field dynamics (2) realize the optimal repertoires of the previous sections when the availability function  $A$  is matched to the cost function  $F(m)$  through the relation

$$A(\tilde{N}_a) = -c' \bar{F}' \left( \frac{\tilde{N}_a}{N_{st}} \right), \quad [3]$$

where  $N_{st}$  is the total number of receptors  $\sum_r N_r$  at steady state. Table 1 shows  $A(\tilde{N})$  for several cost functions. To understand this result, first note that when binding is not cross-reactive, the dynamical equations for each receptor are independent and read  $dN_r/dt = N_r(Q_r A(N_r) - d)$ . The availability function now depends only on  $N_r$ , meaning that receptors compete only with their own kind—they occupy their own antigenic niche. The steady-state size of clone  $r$  is thus set by the carrying capacity of that niche,  $N_r = A^{(-1)}(d/Q_r)$ , or zero if that capacity is negative. With the availability given by Eq. 3, this reproduces the optimal repertoire. As seen in Table 1, fast-growing cost functions correspond to very load-sensitive availability functions. In these cases, rare infections are almost as threatening as frequent ones; therefore the growth of the receptors that are specific to frequent antigens is actively limited to leave room for other receptors. The correspondence of Eq. 3 holds when receptor binding is cross-reactive (SI Text, Appendix J).

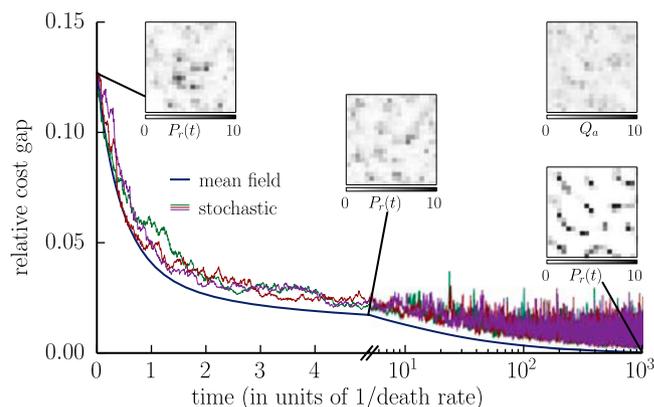
Cross-reactivity leads to competition among receptor types, effectively enforcing an exclusion between similar receptors. This phenomenon, known in ecology as competitive exclusion, is important for lymphocyte dynamics (25) and provides the mechanism by which our dynamical model reproduces the discrete clustering found in the optimal receptor distribution.

To check that the dynamics do converge to the optimum, we simulated numerically the full stochastic dynamics, as well as their mean-field limit (Eq. 2), for a random antigenic environment in two dimensions, with  $A(\tilde{N}) = 1/(1+\tilde{N}/N_0)^2$ . Fig. 5 shows the dynamics of the receptor distribution  $P_r(t) = N_r(t)/\sum_{r'} N_{r'}(t)$ , as well as its cost relative to the optimal solution, as a function of time. Starting from a random distribution of receptors, the repertoire reorganizes into localized peaks that become increasingly prominent and well separated with time. Three independent runs of the stochastic simulation all converge approximately to the global minimum of the cost, with most of the improvement achieved within a few cell lifetimes. Convergence is exact in the mean-field limit, indicating that the steady-state solution is indeed optimal.

In summary, competitive dynamics can allow the immune repertoire to self-organize into a state that confers high protection against infections. In the special case when the availability  $A$  is scale invariant, the expected cost is a Lyapunov function of the dynamics (SI Text, Appendix K), implying that the optimum is reached regardless of the initial condition. Note, however, that the dynamics of Eq. 2 are expected to slow down with age, as the plasticity of the adaptive system decreases due to the diminishing number of naive cells (27).

### Discussion

We introduced a general framework for predicting the optimal composition of the immune repertoire to minimize the cost of infections contracted from a given distribution of antigens. This framework can be extended in several ways to be more biologically



**Fig. 5.** The immune repertoire can self-organize to a state that minimizes cost and provides protection against infections via competitive evolution of receptor populations stimulated by antigens. Numerical simulations of the population dynamics, as well as its mean-field limit (Eq. 2), show how competition causes a random initial receptor distribution to fragment into a highly peaked pattern [insets represent  $P_r(t) = N_r(t)/\sum_{r'} N_{r'}(t)$ ]. *Top Right Inset* represents the antigenic environment  $Q_a$  driving the dynamics [generated from a lognormal noise of power spectrum  $\propto 1/(1+(5q)^2)$  and coefficient of variation 1]. Departure from optimality, as measured by the relative cost gap  $[(F)(P_r(t)) - (F)(P_r^*)]/(F)(P_r^*)$ , decreases with time and eventually reaches zero in the mean-field limit. The three independent runs of the stochastic dynamics show reproducible results. We use the availability function  $A(\tilde{N}) = 1/(1+\tilde{N}/N_0)^2$  with  $N_0 = 10^6$ , a death rate  $d = 0.001$ , and a cost function  $F(m) = 1 - e^{-\beta m}$  with  $\beta = 1/110$ . The space size is  $10\sigma$ . The initial condition was drawn from a lognormal noise of power spectrum  $\propto 1/(1+(5q)^2)$ , with coefficient of variation 2 and  $\sum_r N_r(0) = 1.1 \times 10^8$ . In the stochastic simulations, the time between antigen presentations is  $\Delta t = 0.005 d^{-1}$  (200 infections per cell lifetime).

faithful, e.g., by accounting for antigen-dependent infection dynamics and evolution of the pathogenic landscape. Our predictions can be tested in experiments that study how the environment influences the composition of immune repertoires, either via high-throughput sequencing surveys of receptor populations (2, 28) or by sequencing receptors specific to given antigens (14). The comparison between theory and experiment will provide insight into the functional constraints of antigen recognition by the immune system.

There are many situations where living systems must respond to very diverse and often very high-dimensional spaces of external influences, using strictly limited resources. To sense, internally represent, and then respond to these influences, organisms often use a large diversity of components, such as cell types or genes (29), each sensitive to a small part of the space. For example, the retina supports a diverse population of ganglion cell types, each sensitive to a different visual feature, that collectively represent the behaviorally salient aspects of visual scenes (30). Likewise, the mammalian olfactory system contains some  $\sim 1,000$  distinct receptors that each bind widely to odorants and collectively cover olfactory space (31). In these cases, the limited repertoire of component types provides a key constraint on information processing. Faced with such constraints, living systems must commit resources wisely, adapting to the structure of the environment and balancing breadth of coverage against depth of resolution, in light of priorities, costs, and constraints (32). We have shown that these elements also shape the optimal form of the immune repertoire.

Our finding that cross-reactivity causes the optimal repertoire to fragment is related to the concept of limiting similarity due to

competitive exclusion in ecological settings (33). In this context, empty regions of phenotypic space result when competition is important on the scale at which resources vary, and continuous coexistence of species occurs only in exceptional cases (34). In general, niche-space heterogeneity promotes species clustering (35), recalling our finding that any heterogeneous antigen distribution leads to fragmentation of the optimal repertoire. The conceptual connection between the immune repertoire and ecological organization is even clearer in our dynamical model where species compete for an array of resources (the antigens) and grow in relation to their success in securing resources.

Although this study relies on a simple abstraction of the adaptive immune system, we expect that our framework and results will extend to other distributed protection systems where diverse threats are addressed by an array of specific responses. For example, the immune system of bacteria, or the clustered regularly interspaced short palindromic repeats (CRISPR) system (36), for which population dynamics models have already been proposed (37), could be studied within a similar framework to predict the relative abundance of CRISPR spacers and corresponding viruses in a coevolving population of bacteria and viruses.

**ACKNOWLEDGMENTS.** We thank O. Rivoire and F. Zamponi for helpful discussions. The work was supported by European Research Council Starting Grant 306312. V.B. was supported by the Fondation Pierre-Gilles de Gennes and National Science Foundation (NSF) Grants PHY-1058202 and EF-0928048. Portions of this work were done at the Aspen Center for Physics, supported by NSF Grant PHY-1066293. A.M. was supported by a German Academic Exchange Service (DAAD) Promos stipend.

- Burnet FM (1976) A modification of Jerne's theory of antibody production using the concept of clonal selection. *CA Cancer J Clin* 26(2):119–121.
- Thomas N, et al. (2014) Tracking global changes induced in the CD4 T-cell receptor repertoire by immunization with a complex antigen using short stretches of CDR3 protein sequence. *Bioinformatics* 30(22):3181–3188.
- Forrest S, Perelson A, Allen L, Cherukuri R (1994) Self-nonsel self discrimination in a computer. *Proceedings of the IEEE Symposium on Research in Security and Privacy* (IEEE, Oakland, CA), pp 202–212.
- Perelson AS, Oster GF (1979) Theoretical studies of clonal selection: Minimal antibody repertoire size and reliability of self-non-self discrimination. *J Theor Biol* 81(4):645–670.
- Weinstein JA, Jiang N, White RA, 3rd, Fisher DS, Quake SR (2009) High-throughput sequencing of the zebrafish antibody repertoire. *Science* 324(5928):807–810.
- Mora T, Walczak AM, Bialek W, Callan CG, Jr (2010) Maximum entropy models for antibody diversity. *Proc Natl Acad Sci USA* 107(12):5405–5410.
- Ndifon W, et al. (2012) Chromatin conformation governs T-cell receptor  $J\beta$  gene segment usage. *Proc Natl Acad Sci USA* 109(39):15865–15870.
- Robins HS, et al. (2009) Comprehensive assessment of T-cell receptor  $\beta$ -chain diversity in  $\alpha\beta$  T cells. *Blood* 114(19):4099–4107.
- Larimore K, McCormick MW, Robins HS, Greenberg PD (2012) Shaping of human germline IgH repertoires revealed by deep sequencing. *J Immunol* 189(6):3221–3230.
- Murugan A, Mora T, Walczak AM, Callan CG, Jr (2012) Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proc Natl Acad Sci USA* 109(40):16161–16166.
- Six A, et al. (2013) The past, present, and future of immune repertoire biology - the rise of next-generation repertoire analysis. *Front Immunol* 4:413.
- Antia R, Ganusov VV, Ahmed R (2005) The role of models in understanding CD8+ T-cell memory. *Nat Rev Immunol* 5(2):101–111.
- Regoes RR, Barber DL, Ahmed R, Antia R (2007) Estimation of the rate of killing by cytotoxic T lymphocytes *in vivo*. *Proc Natl Acad Sci USA* 104(5):1599–1603.
- Moon JJ, et al. (2007) Naive CD4<sup>+</sup> T cell frequency varies for different epitopes and predicts repertoire diversity and response magnitude. *Immunity* 27(2):203–213.
- Xia X (1998) How optimized is the translational machinery in *Escherichia coli*, *Salmonella typhimurium* and *Saccharomyces cerevisiae*? *Genetics* 149(1):37–44.
- Press WH (2009) From the Cover: Strong profiling is not mathematically optimal for discovering rare mallefactors. *Proc Natl Acad Sci USA* 106(6):1716–1719.
- Silvestri G, et al. (2003) Nonpathogenic SIV infection of sooty mangabeys is characterized by limited bystander immunopathology despite chronic high-level viremia. *Immunity* 18(3):441–452.
- Chaikin PM, Lubensky TC (2000) *Principles of Condensed Matter Physics* (Cambridge Univ Press, Cambridge, UK), Vol 1.
- Torquato S, Stillinger FH (2003) Local density fluctuations, hyperuniformity, and order metrics. *Phys Rev E Stat Nonlin Soft Matter Phys* 68(4 Pt 1):041113.
- Birnbaum ME, et al. (2014) Deconstructing the peptide-MHC specificity of T cell recognition. *Cell* 157(5):1073–1087.
- Huseby ES, et al. (2005) How the T cell repertoire becomes peptide and MHC specific. *Cell* 122(2):247–260.
- Kosmrlj A, Jha AK, Huseby ES, Kardar M, Chakraborty AK (2008) How the thymus designs antigen-specific and self-tolerant T cell receptor sequences. *Proc Natl Acad Sci USA* 105(43):16671–16676.
- Bhardwaj V, Kumar V, Geysen HM, Sercarz EE (1993) Degenerate recognition of a dissimilar antigenic peptide by myelin basic protein-reactive T cells. Implications for thymic education and autoimmunity. *J Immunol* 151(9):5000–5010.
- Venturi V, Price DA, Douek DC, Davenport MP (2008) The molecular basis for public T-cell responses? *Nat Rev Immunol* 8(3):231–238.
- De Boer RJ, Perelson AS (1994) T cell repertoires and competitive exclusion. *J Theor Biol* 169(4):375–390.
- De Boer RJ, Freitas AA, Perelson AS (2001) Resource competition determines selection of B cell repertoires. *J Theor Biol* 212(3):333–343.
- Miller RA (1996) The aging immune system: Primer and prospectus. *Science* 273(5271):70–74.
- Vollmers C, Sit RV, Weinstein JA, Dekker CL, Quake SR (2013) Genetic measurement of memory B-cell recall using antibody repertoire sequencing. *Proc Natl Acad Sci USA* 110(33):13463–13468.
- Tkačik G, Walczak AM, Bialek W (2009) Optimizing information flow in small genetic networks. *Phys Rev E Stat Nonlin Soft Matter Phys* 80(3 Pt 1):031920.
- Gollisch T, Meister M (2010) Eye smarter than scientists believed: Neural computations in circuits of the retina. *Neuron* 65(2):150–164.
- Buck L, Axel R (1991) A novel multigene family may encode odorant receptors: A molecular basis for odor recognition. *Cell* 65(1):175–187.
- Balasubramanian V, Sterling P (2009) Receptive fields and functional architecture in the retina. *J Physiol* 587(Pt 12):2753–2767.
- MacArthur R, Levins R (1967) The limiting similarity, convergence, and divergence of coexisting species. *Am Nat* 101:377–385.
- Pigolotti S, López C, Hernández-García E (2007) Species clustering in competitive Lotka-Volterra models. *Phys Rev Lett* 98(25):258101.
- Szabó P, Meszéna G (2006) Limiting similarity revisited. *Oikos* 112:612–619.
- Marraffini LA, Sontheimer EJ (2010) CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nat Rev Genet* 11(3):181–190.
- He J, Deem MW (2010) Heterogeneous diversity of spacers within CRISPR (clustered regularly interspaced short palindromic repeats). *Phys Rev Lett* 105(12):128102.