DOI: 10.1111/imr.12664

INVITED REVIEW

WILEY Immunological Reviews

The Bayesian optimist's guide to adaptive immune receptor repertoire analysis

Branden J. Olson | **Frederick A. Matsen IV**

Computational Biology Program, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

Correspondence
Frederick A. Matsen IV, Computational Biology Program, Fred Hutchinson Cancer Research Center, Seattle, WA, USA
Email: matsen@fredhutch.org

Funding information
This work supported by National Institutes of Health grants R01 GM113246, R01 AI120961, and U19 AI117891. The research of Frederick Matsen was supported in part by a Faculty Scholar grant from the Howard Hughes Medical Institute and the Simons Foundation.

Abstract

Probabilistic modeling is fundamental to the statistical analysis of complex data. In addition to forming a coherent description of the data-generating process, probabilistic models enable parameter inference about given datasets. This procedure is well developed in the Bayesian perspective, in which one infers probability distributions describing to what extent various possible parameters agree with the data. In this paper, we motivate and review probabilistic modeling for adaptive immune receptor repertoire data then describe progress and prospects for future work, from germline haplotyping to adaptive immune system deployment across tissues. The relevant quantities in immune sequence analysis include not only continuous parameters such as gene use frequency but also discrete objects such as B-cell clusters and lineages. Throughout this review, we unravel the many opportunities for probabilistic modeling in adaptive immune receptor analysis, including settings for which the Bayesian approach holds substantial promise (especially if one is optimistic about new computational methods). From our perspective, the greatest prospects for progress in probabilistic modeling for repertoires concern ancestral sequence estimation for B-cell receptor lineages, including uncertainty from germline genotype, rearrangement, and lineage development.

KEYWORDS
Bayesian inference, high-throughput sequencing, likelihood models, repertoire analysis

1 | INTRODUCTION

1.1 | Why bother with probabilistic models?

Before entering on our quest for model-based analysis of repertoires, one might ask "why bother?"

The first answer is simple: repertoires are generated by a probabilistic process of random recombination, unknown pathogen exposures, and stochastic clonal expansion. Thus, when analyzing repertoires it behooves us to reason under uncertainty. The last century of statistical development offers a refined set of tools to make statements about such systems and assess our confidence in them.

Second, repertoire data show us that complex models are justified. For example, not all germline genes are used with equal frequency in repertoire generation. The frequency of these germline genes is interesting to measure, but also informative of which genes were used in specific recombination events that gave rise to observed sequences. Furthermore, the various genes all have characteristic distributions of trimming lengths, shown to be consistent between individuals[4–6]; incorporating this further improves annotation and clustering inference. Such observations can also suggest mechanistic hypotheses that can then be tested with experiments.

Third, the probabilistic approach offers a principled means of accounting for hidden latent variables that form an essential part of the model, but are not themselves of direct interest to the researcher.

This article is part of a series of reviews covering Characterization of the Immunologic Repertoire appearing in Volume 284 of *Immunological Reviews*.

For example, we may not care about the exact rearrangement event that led to a given B-cell receptor, but this is still an important latent variable for clustering analysis: indeed, one should only cluster receptors that came from identical rearrangement events. Thus one can sum over the possible rearrangement events that led to this clonal family, leading to a natural means of evaluating a clustering likelihood [7] that averages out uncertainty in the rearrangement process.

Fourth, probabilistic models have well-developed notions of model hierarchy, in which inferences at each level inform and are informed by inferences at other levels. This is essential to leverage the hierarchical structure present in immune receptor data (Figure 1). For example, performing inference using many sequences at once (eg, germline inference) can greatly improve per-sequence inferences, performing lots of per-individual germline inferences can tell us about the germline biases of a population, and so on up the hierarchy.

1.2 | Model-based probabilistic analysis

We begin by introducing model-based probabilistic analysis, and providing a very casual introduction to maximum likelihood and Bayesian analysis as they apply to immune repertoires.

Consider a very simple model of the distribution of heights in a human population: a normal (a.k.a. Gaussian) distribution. Say we have observed the height of all 127 million humans in Japan, rounded to the nearest centimeter, and we have plotted it as a histogram. As a first approximation, one can think of fitting a probabilistic model as grabbing a normal distribution and flexing it with our hands until it looks as much as possible like that histogram. If its estimates are too small, for example, we can scoot it right, and if it is too narrow we can bend it so it is broader.

This process can be formalized in terms of the principle of maximum likelihood, in which we find the parameter values that are most likely to have generated the observed data. The likelihood function of the model parameters can again be thought of as "the probability of obtaining the observed data under the given model with those parameters." Although not quite a rigorous definition for all settings, this definition is rigorous for discrete data such as heights rounded to the nearest centimeter, or DNA sequences. For a normal distribution model, which is parameterized by mean μ and variance σ^2, we can directly calculate this likelihood function. This likelihood is a product of terms, one for each human, equal to the Gaussian probability density $(2\pi\sigma^2)^{-1/2} \exp[-(x-\mu)^2/2\sigma^2]$, where x is the height of that human. It turns out that the usual formulas for the mean (ie, the sample average) and the (biased) sample variance (the average squared deviation from this mean) are exactly the maximum likelihood values of these parameters: those values that maximize the likelihood function!

One can also approximate the likelihood function using repeated simulation for a single set of parameters, as we illustrate using the following thought experiment. In the heights example, we can estimate the likelihood as follows: generate many samples of size 127 million from a normal distribution rounded to the nearest integer, and calculate the fraction of times we get exactly the observed set of heights. Although this will be an extraordinarily small number, it will be larger for parameter values that fit well (the normal distribution fits the histogram of measurements closely) than for ones where it does not, and thus is a means of doing parameter fitting. [Note that these two perspectives on optimizing our model, that of picking model values such that simulation is as close as possible to observation, and that of maximizing a likelihood function, are actually identical if we define "as close as possible" in terms of Kullback-Liebler [8] divergence.]

The inferential setup is the same for immune repertoire analysis, except that the models and data are more complex (Figure 2). Rather than having a model that generates human heights, models for immune repertoire analysis generate immune repertoires as collections of DNA sequences. In a similar way, fitting such models is a process of wiggling the parameters until the model generates repertoires that are as similar as possible to the observed repertoires. In certain cases we can efficiently compute a likelihood function such that optimizing this function does the wiggling more formally, using either an exact formula or approximate numerical routines. However, this is not the case for all models, and indeed much of the subject of the

How to read this paper

- If you are an immunologist and want to learn more about probabilistic modeling, start here.
- If you love probabilistic modeling and are curious about immune repertoires, you may want to start by getting background in immunology in general [1] and immune repertoires in particular,[2,3] then reading the Models section.
- If you already know both topics and get bored easily, skip to your favorite parts of repertoire analysis.

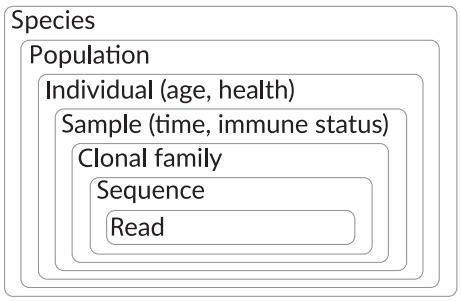

FIGURE 1 Immune repertoires are hierarchically structured, here illustrated by the hierarchy for B-cell receptor sequences. We benefit by considering the whole hierarchy that contributes to our observable sequences rather than one sequence at a time. For example, by considering all the reads at once one can infer a personal germline set, which then informs the per-read annotation. By learning lots of personal germline sets one can infer population-level germline trends

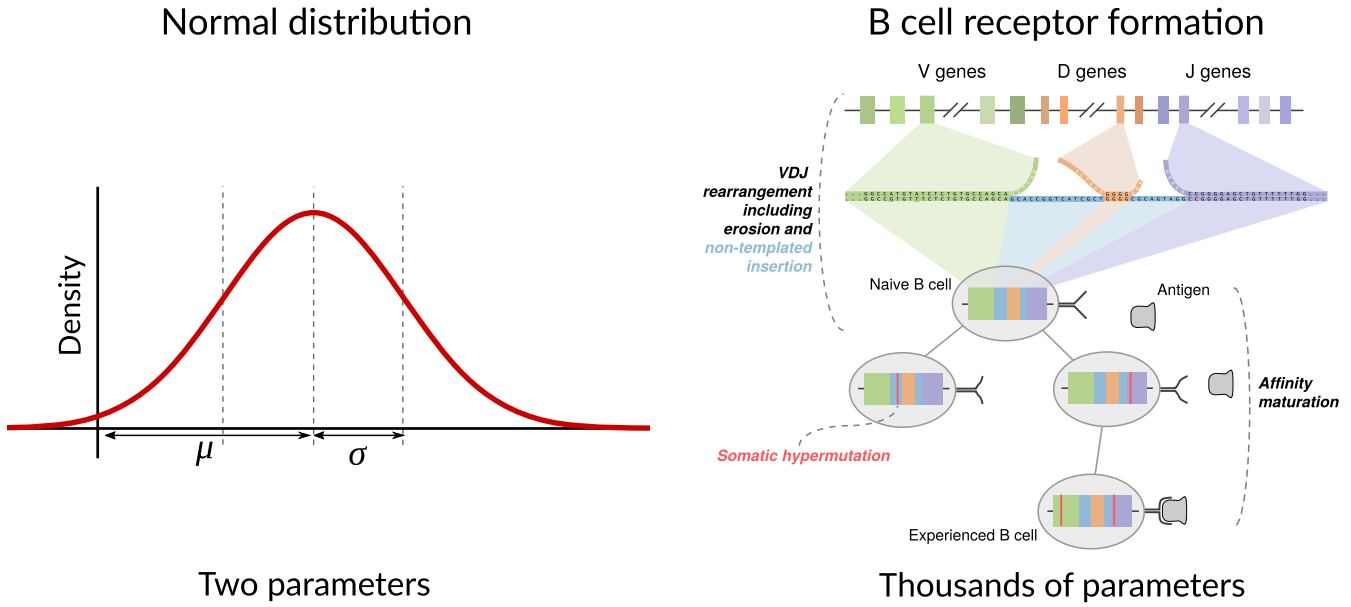

FIGURE 2 Human height and immune receptor formation can both be modeled using probabilistic methods. Right panel modified (with permission) from (4)

second half of this paper describes models that attempt to make a balance between computability and realism.

For repertoires, we can again imagine maximum likelihood fitting happening via simulation: we have a model from which we can simulate repertoire sequences, and we can approximate the likelihood for a collection of parameters for a given dataset based on the fraction of times it generates the observed data exactly. In principle, we can fit the model by iteratively wiggling parameters and re-simulating, preferring those wiggles that more frequently generate the same data as what was observed. Of course, for real datasets such fitting is sheer lunacy: a repertoire simulation will never once exactly match an observed repertoire sample containing a million unique sequences even if we were to run it for our whole lifetimes! Nevertheless, this is a helpful thought experiment that underlines the importance of likelihood functions, which can be thought of as a "short cut" avoiding such simulation. We will continue this thought experiment below.

We can continue the height metaphor to explain Bayesian analysis. Bayesian analysis again concerns model parameters θ. In the heights example, θ is a pair consisting of the mean μ and the variance σ^2. The goal of Bayesian analysis is to not just find the best parameters θ, but to get an ensemble of possible values of the parameters along with an idea of how well each describes the data x. This is formalized in the notion of a posterior distribution, which is a probability distribution on the collection of parameters describing how likely the various parameters are to be correct given the data. Having a full distribution over parameters rather than just point estimates allows for a more detailed characterization of the uncertainty in our inferences. For example, we can summarize this posterior distribution in terms of credible intervals, which are the Bayesian analog of confidence intervals. To obtain such a posterior distribution for our height example, we begin by specifying a prior distribution on the parameters. This prior is our a priori idea of what the heights might be before we sample any data. We then incorporate the data to get a posterior distribution. Formally, this comes from the deceptively simple statement of Bayes' theorem:

$$p(\theta|x) \propto p(x|\theta)\, p(\theta)$$

which states that the posterior distribution $p(\theta|x)$ of model parameters θ given data x is proportional to the likelihood $p(x|\theta)$ times the prior $p(\theta)$. We can think of this as re-weighting our prior assumptions based on how well they explain the data.

This sounds simple enough, but in fact thousands of careers of computational Bayesians have been dedicated to the challenge posed by Bayes' theorem being expressed in terms of proportionality rather than equality. Indeed, even if we can say how much one parameter set is better than another via Bayes' theorem, we have to evaluate many different parameters to obtain a value for the posterior, which makes a statement about how good a given parameter set is compared to *all* possible parameters. The situation is analogous to that of climbers in a mountain range tasked with estimating their height relative to the average height of the range: it is easy to see that one location is higher than the other, but evaluating the average height requires traversing the entire range and taking careful measurements. This is an informal way of saying that the integral of the posterior distribution is typically intractable.

When the posterior integral is in fact tractable, as can be the case for very simple models, we can obtain the posterior distribution directly as a formula. In our height example, if we take a normal prior distribution for the mean with a fixed variance, we can directly obtain a formula for the posterior distribution (which turns out to also be normal). However, such directly computable models with

so-called "conjugate priors" are few and far-between, and none of them involve immune receptor DNA sequences.

For more complex models we do not attempt to compute the posterior distribution directly, but rather we sample from it. In this way we obtain a "histogram" that approximates the full posterior distribution: in our heights example, we would get a collection of (μ, σ^2) samples from the joint distribution on these parameters. It is common to summarize these samples in terms of their single-variable posterior estimates, which in our example would be one histogram for the mean of the height distribution and another for the variance.

Although sampling from posterior distributions is a challenging problem, decades of research has developed sophisticated methods, as well as probabilistic programming languages that are dedicated to the task.[9–11] We will briefly summarize one method Markov chain Monte Carlo (MCMC) below, but first present a completely rigorous but utterly impractical means of sampling from a posterior distribution via simulation. In the above thought experiment, we were approximating the value of the likelihood for a single set of parameters, and here we have an even more ambitious goal: to approximate the posterior across parameter values. Repeat the following process to obtain a posterior sample on parameters given some data:

- Draw values of the parameters from the prior
- Simulate data using those parameters
- Does this simulated data match the observed data exactly?
- If so, add these parameters to our posterior sample, and if not discard them
- Return to the first step until the desired number of samples is obtained

In the height example, each such cycle involves drawing 127 million samples from a normal distribution and checking if they are the same as the observed data. The result is a sample from the posterior distribution on μ and σ^2. In an immune repertoire example, we could do the same by simulating sequences, which is even less practical than the completely impractical idea of applying this to the height example. Luckily, there are other means of sampling posterior distributions.

The most common method for sampling from a posterior distribution is MCMC. MCMC is a random procedure that moves around parameter space such that the frequency with which the procedure visits a given parameter is proportional to its posterior probability. The most popular type of such inference in phylogenetics is random-walk MCMC,[12] in which parameter values (such as a tree topology and its branch lengths) are perturbed randomly; these perturbed values are always accepted if they are "better" and accepted with some probability if they are "worse." Being able to accept "worse" parameter modifications is important so that the algorithm explores the entire space rather than getting stuck at the peak of a distribution. The notions of "better" and "worse" are determined by the Metropolis-Hastings ratio, which depends on having a likelihood function that can be evaluated efficiently. This sort of sampling is implemented in packages such as BEAST[13] and MrBayes,[14] but due to computational complexity is typically limited to hundreds of sequences in a single tree.

Before exploring computational challenges, we describe marginalization and discuss priors. Marginalization is the practice of "integrating out" nuisance parameters, which are parameters that are important for the model but may not be of interest for the researcher. Imagine we were interested in what D gene was used for a given B-cell receptor sequence, and want to take a probabilistic approach because such assignment is naturally uncertain. In a likelihood-based approach, one can only evaluate the suitability of a D gene assignment when we also have specified the amount of trimming encountered by this D gene, even if that parameter is not actually of interest to us. Therefore we sum over the possible amounts of D gene trimming. In general this is called integration because summation is a special case of integration.

Prior distributions require careful consideration. All distributions, including prior distributions, have parameters that must be chosen. The parameters of prior distributions are called "hyperparameters." Where do those come from? One option is to use a hierarchical Bayesian analysis in which we consider prior parameters as random variables themselves, also requiring prior distributions. The phylodynamics community have developed sophisticated methods to infer mechanisms of viral spread using such a hierarchical approach.[15] However, at some point this recursion must end and one must either fix values arbitrarily or attempt to estimate them from the data. The process of estimating fixed hyperparameters is known as empirical Bayes.[16]

1.3 | An informally described hierarchy of inferential difficulty

Here we describe a difficulty hierarchy for maximum likelihood and Bayesian inference based on how difficult the model is to compute.

1. *Conjugate priors available for model*: In this case, the posterior is available as an exact formula. Hence, no sampling is required, and the posterior is extremely efficient to evaluate. Unfortunately, this is never the case for repertoires.
2. *Efficiently computable likelihood function available*: Here, maximum likelihood estimation is tractable, and Bayesian methods can be used via MCMC. Phylogenetic trees under models where each site evolves independently fall into this category, as the Felsenstein algorithm[17] provides a means for efficient likelihood evaluation. Nevertheless, tree inference is still challenging, and provably hard (in the technical sense) given difficult data[18] because of the super-exponential number of trees that must be tried in order to be sure of finding the best one. Repertoire analysis methods such as hidden Markov models (HMMs, described in more detail below) for rearrangement inference also fall into this category. In this case there is also latent state (ie, the transition points between germline sequences and the N/P junction between germline-encoded regions); this latent state can be efficiently marginalized by the

Forward-Backward algorithm.[4,19] Bayesian estimation for such parameters is also possible[20] though has not been applied to repertoires.

3. *Efficiently computed likelihood function available if we condition on some additional latent state*: Some models do not have an efficiently computable likelihood function in general, though a likelihood can be computed if we expand the parameters of interest to include some additional information. For example, the ideal phylogenetic reconstruction method for repertoire data would take the context-sensitive nature of somatic hypermutation into account.[21] We can efficiently compute a likelihood function using a context-sensitive model such as S5F [22] if we specify the order of and time between mutations. However, these additional parameters are not typically of interest and thus need to be marginalized out using Markov chain methods.[23] For certain classes of such models, only the order of mutations (vs their exact timing) matters.[24]
4. *No likelihood function available*: When no likelihood function is available one must resort to simulation-based methods such as approximate Bayesian computation (ABC).[25] In this method, one obtains approximate posterior distributions by reducing the data to relevant summaries and seeing which models produce data that match these summaries well. Our above thought experiment required an exact match of simulated and experimentally derived data in order for a set of parameters to be accepted. In ABC, one accepts parameters with a probability determined by how closely prespecified summary statistics of the simulated data agree with those of the experimental datasets. This has been applied with success in population genetics problems with a modest number of parameters. However, as the model complexity grows, even simulation-based methods suffer the "curse of dimensionality" and will eventually become intractable.

Any sufficiently detailed model of repertoire generation will land here. For example, it is not possible to calculate likelihoods for complex models based on agent-based simulation,[26] although one could sample them using ABC. In fact, an informal version of ABC is currently used in B-cell receptor sequence analysis, in which one adjusts simulation parameters until they generate data that looks close to experimental data according to a battery of summary statistics.[27–29]

We see that there is often a balance between realism and computability; although there is no inherent reason why this must be so, it is often the case. For example, computation is eased by assuming variables in a model are independent, even if that is not exactly true. In the above hierarchy, this is illustrated by easy-to-compute site-independent phylogenetic models on one hand vs hard-to-compute context-dependent models on the other.

2 | MODELS

Here we describe existing and potential probabilistic models for immune receptor development. Although in principle any probabilistic model can be used for inference (via the "thought experiment" inference procedure described above), we find it useful to distinguish between inferential models and models for simulation. For the purposes of this paper, inferential models are those that are meant to be fit to data to learn something about the underlying system.

We will be interested in inferential models that are tractable to use for infer. ence if one is "optimistic" (marked with ☼): at least, one should be able to do inference on each individual component using existing machinery.

Models for simulation serve a separate and essential purpose. Such models can be more complex and need not have an efficiently computable likelihood to be useful. Agent-based models, such as models of a germinal center[26] fall into this category. Models can make predictions, such as the groundbreaking 1993 prediction of cyclic re-entry[30] that was dramatically validated over a decade later.[31,32] Also, if we want to validate inferential algorithms, we need accurate generative models. For these reasons we are going to sketch "lunatic" model components (marked with ☾) as well, for which we only require the ability to simulate in forward time.

We will investigate this framework while following receptor development from the germline gene repertoire to clonal expansion. For every component of the process, we will follow an identical pattern in this order: biological background, then previous work on inference, then sections on "optimist" ☼ and "lunatic" ☾ models. The biological background will of course be a miniscule fraction of what is known, as we can only include parts that are relevant for the modeling goals here.

Before we begin this voyage, we note that traditionally biologists and statisticians have slightly different but not incompatible notions of what is meant by "model." A biologist's model is typically a conceptual model describing the mechanistic process by which something happens. For example, transcription factor *X* binds cofactor *Y* which allows it to initiate transcription of gene *Z*. Such a model may not have any parameters and thus cannot "generate" data, although it can typically be used to devise an experiment to test the hypotheses of the model.

A statistician's model, on the other hand, need not have any mechanistic underpinning, although it necessarily contains parameters and can be used to generate data.

These two perspectives lead to different means of iterative model improvement (Figure 3). Biologists scrutinize their models for components that can be separated out and perturbed individually to form a test of the model. This reductionist approach has taught us most of what we know about biology today. Statisticians, on the other hand, are generally interested in evaluating models via model fit. That is, if we generate data from our model, does it resemble our observed data, and are new, unseen data values described well by the model? If not, how can we add model components that will result in a better-fitting model? This iterative process of model improvement has been called "Box's loop."[33]

Nonetheless, these viewpoints are quite compatible, and indeed we may need to combine them to meet the next set of challenges in adaptive immune receptor research. For the statistician, incorporating mechanism into statistical models means that inferred

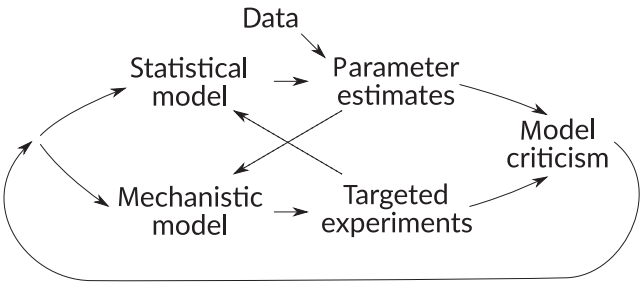

FIGURE 3 Statisticians and biologists have typically approached adaptive immune receptor research in two parallel tracks. Statisticians (upper path) treat data as given and perform model criticism based on numerical estimates of how well models fit the data. Biologists (lower path) formulate their models mechanistically and generate data in targeted experiments to directly test their model. An integrated approach (dashed arrows) has biologists using parameter estimates to formulate mechanistic models, and statisticians using the results of targeted experiments to formulate statistical models. Ideally, the distinction between these two classes of models would evaporate, although mechanistic models are not always readily fit using statistical means

parameters have direct interpretation, and such models typically have fewer parameters. For the biologist, formalizing a biological model statistically means that models with hidden parameters can be directly compared in a rigorous way.

With all this introduction out of the way, now we begin considering probabilistic models for adaptive immune repertoires!

2.1 | Germline genotype

Although repertoire modeling often starts with V(D)J rearrangement,[34] the rearrangement process is in turn determined by the genotype of each individual, in particular the collection of germline V, D, and J genes in the loci forming the various receptors. A complete survey [35,36] is out of the scope of the paper, but suffice it to say that although complete haplotypes at germline loci have been sequenced,[37–40] the genetic diversity of these loci is high and many new alleles continue to be discovered, especially in non-Caucasian populations.[41–44] Thus, the germline genotype forms an important part of the hidden state for repertoire generation[45] with real medical consequences.[41,46]

This motivates inference of germline genotypes directly from repertoire sequence data. Early work performed such inference by carefully considering several sets of high-throughput sequencing data.[5,47,48] Kidd et al[48] used a maximum likelihood model assuming uniform gene use to infer alleles, and were able to phase these onto haplotypes for individuals who are heterozygous at *IGHJ6*. Later work used naive sequences and an assumption of no gene duplication to iteratively obtain haplotypes via probabilistic gene assignment for three individuals.[5] More recent work has delivered automated tools for germline set inference: TIgGER,[43] IgDiscover,[49] and partis.[50] TIgGER introduced a "mutation accumulation" plot relating mutations at a given site to the overall level of mutation in sequences. This plot should have a smooth shape in the absence of new alleles, but a "bend" in the presence of an unannotated allele; partis works by explicitly searching for this bend. IgDiscover applies hierarchical clustering to naive-sorted data to obtain germline sets in species for which little or no germline information is known.

☼ There is a substantial need for probabilistic germline repertoire inference methods. For example, this would be very useful if we want to estimate the unmutated ancestor of B-cell clonal families while quantifying uncertainty.

Germline gene inference is inferred from a whole repertoire at a time; this fact alone poses some challenges to a probabilistic method. For example, if we wish to compare two germline gene sets, naively one would need to perform a complete re-alignment of all sequences to obtain a likelihood. Because such a likelihood is available, it can be formally classified as "efficient" (second category in the above hierarchy) although such repeated re-alignment is not practical for large datasets. Some cleverness would be helpful here, such as only re-annotating sequences that could be affected by changing the germline set. Another alternative would be to infer a too-large pool of possible candidate germline genes, perform probabilistic alignment using all of these germline sequences, and use the associated probabilities without running complete realignment in a second step to cut down the pool.

Haplotype inference has been shown to be a useful tool for cutting down germline sets from such a candidate pool.[5,51] This works by assuming limited or no gene duplication of individual genes in a germline set, and then using joint gene usage under VDJ recombination to infer which alleles lie on which haplotype. One can then review the gene assignments and reject suspicious inferences, since having many alleles of a given gene inferred to lie on a single haplotype casts doubt on their authenticity. This method does have some caveats. It requires heterozygosity at J (or D) genes, and that the V gene in question is expressed at reasonable levels on both chromosomes. Also, the immunoglobulin germline locus is dynamic with many gene duplication and conversion events, so we cannot exclude the possibility of many alleles of a gene being present on a haplotype.

Despite these caveats, in order to have the best germline gene inference it may be useful to extend inference to the full pair of haplotypes, called a "diplotype." In order to do so we will need to formalize generative models on diplotypes, which could act as a prior for inference. Ideally, such a generative model would come from observing many diplotypes. As noted above, such direct haplotype sequencing is rare, but this situation may change using improved assembly techniques applied to long-read sequencing data. Alternatively, one could build up such a model by taking a large ensemble of datasets and iteratively estimating haplotypes and prior parameters (determining, eg, the prior distribution of the number of alleles per gene) using empirical Bayes. A parameterized prior could be developed based on racial background, in which people with genetic ancestry from various places would have a different distribution of germline genes. The biological importance of gaining broad diplotype information has been carefully laid out in a recent review.[45]

In principle one could directly use probabilistic methods to infer a pool of possible germline sequences, although here the problem of requiring reannotation becomes much more acute because of the many hypotheses that must be tried. The TIgGER mutation accumulation plot gives the values of a complex conditional probability, and at least for the near term it makes sense to continue using this summary. Its analysis could be improved by more flexible models of how mutations accumulate on sequences—current efforts implicitly assume that each site accumulates mutations linearly as a function of the total number of mutations on that sequence.

☾ Germline genotypes differ between individuals in a population and between populations and species because of long-time-scale evolutionary processes; it is already tempting to work to better understand these processes of mutation and selection. In terms of mutation, one may wish to connect germline gene change with mechanistic models of gene duplication and loss.[52] Analysis of large populations,[53] especially using parent-offspring data,[54] will be important to develop such models. It is also tempting to infer selection on germline gene sets to maintain a diverse pool of starting material for VDJ rearrangement, as well as to make it easier to mount an antibody-mediated response against locally important pathogens. Much more data, in particular data spread across many more populations, will be required to perform such inference.

2.2 | Rearrangement

By "rearrangement" we mean joint gene choice, trimming, and insertions in the process of V(D)J recombination [55] without any selective steps for tolerance or binding. Biologically, this is determined at least in part by gene location, presence of recombination signal sequences,[56] chromatin accessibility of these sequences,[55] and long-range loop structure.[57] There are surely additional complex genetic determinants of the rearrangement process, and how those contribute to repertoire formation will be a continued topic of research.

This complex machinery leads to a complex probabilistic process that determines rearrangement.[36] Prior work has found interaction between N nucleotide addition and recombination [58] as well as dependence between D and J gene use in BCRs.[59] There is also clear evidence of interaction between gene use and trimming length.[4,60] The rearrangement processes change with age, a phenomenon recently quantified in mouse.[61] In addition to the usual rearrangement process, oddities such as VH replacement [62–64] and inverted and multiple D genes do occur, although recent analyses indicate that these are rare in the overall repertoire.[65,66]

This process is greatly deserving of complex models, as all variables determining the rearrangement process are both interesting and decidedly non-uniform. Indeed, many probabilistic models have been formulated, with the hidden Markov model (HMM) framework being particularly popular.[6,60,65,67,68] Various implementations of the HMM differ in the parameterization of gene choice and trimming distributions, with the trend being toward parameter-rich categorical distributions for trimming. Such rich distributions are justified by the observation that although trimming distributions are different between genes, strong concordance between individuals shows that the models are not simply fitting noise.[6,60,69] Recent work has extended this to a more general modeling framework expressible in terms of an arbitrary Bayesian network.[69] For the insertion sequences, applying an HMM has shown dependence of the next base on the previous one.[60]

☼ Despite substantial progress, there is still work to be done describing the rearrangement process using probabilistic models. The distribution of inserted sequences invites further exploration: are more complex models warranted? Although current models are inferred per-data-set, it would be helpful to have models that can concern multiple related datasets and parameterize differences between them using covariates such as age.[61] Such models may also be useful to infer differences in the rearrangement process by genetic back-ground.[41,46,70] An obvious if formidable next step is to extend current probabilistic models to include complex rearrangements such as replacement and multiple D genes.

☾ It will be more challenging to relate these descriptive statistical models to mechanism. As described above, gene choice is determined by recombination signal sequence strength and accessibility. Sequence features must also govern the amount of trimming, and early work found sequence motifs that change the distribution of trimming amounts.[71,72] This work has not been extended in our current era of abundant high-throughput sequencing datasets. However, our biological knowledge has also expanded: we now know gene choice is determined by processes including megabase-scale loops and chromatin state, although the roles of various processes such as hairpin opening and nucleotide deletion to germline gene trimming are still something of a mystery. Since these processes are so complex, any proposed model will have to judiciously choose a balance between realism and tractability. Also, such a project will require diverse expertise in statistical modeling and biological mechanism.

2.3 | Initial selective filters

Positive and negative selection determines which B and T cells are able to circulate. Positive selection ensures that T cells are able to bind major histocompatibility complex (MHC) molecules of the host. Negative selection happens to avoid self-reactivity and maintain an appropriate level of interaction with MHC. In B cells the initial selective processes ensure that a functional antibody is produced with limited self-reactivity.

For B cells, previous work has found selection against long and/or hydrophobic HCDR3 loops[73] as well as selection on germline gene use[74] and D gene frame.[75] Others have inferred "selection factors" for various aspects of the TCR[76] and BCR[5] in the initial selective process. These factors are multiplicative terms that describe the probability of seeing a sequence with a specific characteristic in the post vs preselection repertoire. For example, a selection factor for a specific amino acid at a specific location quantifies the level to which this amino acid is selected for or against.

☼ There are still many possible ways to extend analysis of "bulk" characteristics of sequence-level selection. Current methods

analyze the role of a single feature at a time, such as CDR3 length or particular selection factors, and it would be interesting to look for selection on sets of factors. Paired heavy/light and alpha/beta chain data is also an interesting source for such joint selection analysis.[77–79] It will also be important to take MHC type into account when performing such analysis and look for MHC-mediated effects, although the largest analysis so far found relatively few TCRs that were negatively associated with MHC.[80]

☾ It will be very difficult to make the leap from such a "bulk" analysis of sequence-level selection to the true prize, which is inference on a per-sequence level. Extracting the binding properties of a BCR or TCR from sequences, and hence its potential for autoreactivity, is a grand challenge of computational biology that will not be solved soon. Perhaps the best strategy will be via protein structural modeling, or via machine learning techniques applied to large datasets of pre and postselection receptor sequences. Thus such per-sequence analysis appears to be out of scope of the sort of probabilistic modeling considered here.

2.4 | T-cell clonal expansion

T cells are stimulated to divide when they bind to an MHC loaded with a peptide that they recognize. This is called clonal expansion.

As in the previous section, one can again consider two questions for clonal expansion: first, what are bulk characteristics of the expanded repertoire, and second, can we infer anything on individual TCR sequences? Regarding bulk characteristics, abundance distributions of T cells have proven to be a fertile means of learning about the patterns of antigenic stimulus and competition.[81,82] Twin studies show a strong genetic effect of T-cell clonal expansion in terms of overall memory cell response and response against a specific immunization.[83] T-cell development changes through age, which has been used to show that our naive T-cell repertoire is a complex mixture of cells generated at different ages.[61,84] Aging clearly modifies both the existing repertoire and our capacity to respond to novel stimulus.[85]

Regarding individual sequences, it is interesting but difficult to associate characteristics of specific TCR sequences with genetics and immune state. One component of this is to develop relevant notions of similarity between receptors, which can then be used to perform clustering and projection into a lower dimensional space.[86,87] Using these and related tools, recent work has moved toward a variety of machine learning goals, including clustering sequences according to their specificity using tetramer-binding data,[86,88] predicting new sequences that will bind a given epitope,[88] identifying relationships between TCR sequence and MHC use,[80,89] and finding sequences or sequence characteristics that differ between groups.[80,90–92] Building databases of epitope-TCR pairs[93] and high-throughput measurement of affinity[94] will certainly spur this development.

☼ Probabilistic modeling can be used to estimate the chance of obtaining a given TCR sequence, and thus has an important role in interpreting T-cell frequency data. Such models can be used to estimate the degree of antigen-stimulated clonal expansion by comparing the probability of generation to the TCR frequency.[92] In addition, probabilistic modeling is appropriate for interpreting bulk properties of repertoires in terms of clonal relationships within the sequences. Here, we will certainly see continued development of models describing the abundance distribution of T-cell clones, and how these change with age and immune stimulus.

On the other hand, predicting biophysical properties of individual TCR sequences is not easily solved using a model-based probabilistic framework. The binding fitness landscape of individual sequences is too "rough" for typical probabilistic methods—a small modification in the right place can take a strongly binding receptor and make it non-binding.

☾ Thinking about a generative model that is too complex for inference, we seem to fall between two stools: probabilistic generation models and models of bulk properties seem tractable, whereas models of individual sequences seem impossible or inappropriate with current probabilistic tools.

2.5 | B-cell clonal expansion: antigenic stimulus

B-cell clonal expansion is complex and delicious for statistical modeling, and thus we will divide our description of this process among the following five sections.

In response to immune challenge, dense accretions of lymphocytes form structures known as germinal centers.[95] These are the sites of B-cell diversification, to which B cells gain entry by the ability to bind antigen. This diversification process includes mutation and selection processes that will be covered in subsequent sections below. After responding to the original infection, memory B cells and plasma cells are exported from the germinal center; after export these cells are mostly dormant until further stimulated. Serum antibody levels are maintained by dynamically controlled cell populations.[96]

Methods are now emerging to predict antigen-antibody affinity from sequence data for specific antigens.[97] One approach is to find shared sequence characteristics. For example, convergent sequence characteristics such as gene usage, CDR3 length, and mutation load have been identified in response to some vaccines.[98] However, not all sequence characteristics will have straightforward correlations: for example, an influenza vaccination experiment with closely sampled time points did not see a correlation between VJ gene usage and degree of expansion.[99] Age is an important covariate of this sort of analysis, as it changes the degree of hypermutation,[100] how frequently certain gene combinations are generated in the unselected repertoire, and how gene combinations are selected in the memory repertoire.[101]

Another approach is to build out databases of BCR sequences responsive to specific antigens and look for similarity between them, in terms of both sequence similarity and time dynamics of re-activation.[102–104] There is some, though not plentiful, data with which to infer these patterns, and the most information is available for HIV antigens.[105,106] This sort of approach may be aided by

improved modeling of the space of antigen-binding sequences, such as with maximum-entropy models,[107] or some other means of predicting binding similarity using sequence information.

The effectiveness of a database-matching approach depends on similar BCR sequences being used to bind a given antigen, which leads to the subject of "public" repertoire analysis. This sort of analysis determines which sequences are shared between individuals due to common antigen exposure and relatively high-probability random sequence generation.[108–110] To make sense of the public repertoire one must understand the extent to which genetics and negative selection determine the naive repertoire. For example, vaccination of human twins gives rather different results,[111] whereas there is significant evidence of genetic predetermination for vaccination in mouse.[112] Indeed, it has recently been proposed that this represents a fundamental difference between the immune systems of these two species.[113]

One can also consider various bulk properties of the memory BCR repertoire, and consider the difference between the naive repertoire and the mature repertoire. A deep sequencing study observed differences in gene usage and CDR3 length.[114] In addition, using appropriate laboratory and computational strategies, one can quantify and model the respective abundance distributions.[114,115]

.☼ Like TCR sequences, probabilistic models are needed in public repertoire analysis of BCR sequences to disentangle the roles of similar rearrangements and antigenic stimulus in generating similar or identical receptors. Analogously, models of abundance distribution should inform us about the dynamics of generation and selection, although there does not appear to be much work in this area yet. Representations of sequences such as maximum-entropy models may provide useful tools for characterizing groups of antibody sequences binding a given antigen.

Although methods with more of a machine learning flavor may be a better fit for inferences on individual sequences, it may be useful to combine probabilistic models of sequence generation with probabilistic models of sequence families binding a certain antigen.

☾ As for T-cell receptors, we again fall between two stools.

2.6 | B-cell clonal expansion: somatic hypermutation

When B cells replicate, their BCR locus is mutated at a rate about a million times higher than in normal replicating cells. This process is orchestrated by a complex set of steps, starting with deamination of a cytosine to make a uracil, and then proceeding down one of multiple paths of error-prone repair.[116] These steps lead to complex context dependence, determining which antibodies are reachable via somatic hypermutation.[117]

Several decades of work has focused on how these context "motifs" change mutability, first by finding "hotspot" motifs that are especially mutable[21,118,119] and then later by developing more quantitative approaches to describe the influence of various sequence characteristics. This includes models estimating the mutability of all possible sub-sequences of some length[22,120,121] or models that use sequence position and/or presence of individual bases at specific distances to predict mutability.[5,122] Our group has recently generalized these approaches into a penalized survival analysis framework that can combine arbitrary sequence features, omitting those which do not clearly contribute to improved model fit.[24] An alternative way to formulate somatic hypermutation is to consider substitution frequencies of the combined mutation and selection processes on germline genes,[6,123–125] although this does not provide predictions for N-region nucleotides.

Accurate mutation rate estimation is important for interpretation and prediction of B-cell evolutionary patterns. This is clearly true for estimation of natural selection,[126–128] in which the mutation rate (the rate of introduction of nucleotide changes) is compared to the substitution rate (the rate of such changes that persist in the population) in order to estimate a natural selection parameter. It is also important for understanding which antibodies are accessible from a certain rearrangement[117]; analysis of B-cell evolution over long timescales suggests decreasing mutability through time.[129]

Currently there is an unfortunate division between biologically based mechanistic models and statistical models, which are so far only descriptive. Although the pattern of mutations has been used to state qualitatively that certain factors are important in the SHM process,[118,130–133] this has not resulted in rate estimates for the various repair pathways. The one exception is a mathematical model of AID activity in terms of scanning and catalysis,[134] although other processes are essential to the somatic hypermutation process in vivo.[135] A more mechanistically explicit model of DNA damage and local error-prone repair should generate locally correlated sets of mutations more effectively, a task at which current models fail.[69]

In addition to the point mutations described above, somatic hypermutation also introduces insertion-deletion mutations, or indels.[136] The rate of indel introduction is comparable to the rate of point mutation[137–139] although most of these indels are filtered out by natural selection in the functional repertoire. Because the mechanisms for point and indel mutation are linked,[116] it is perhaps not surprising that correlation can be found between their locations.[139] Although most indels are filtered out, some have important functional consequences, such as in the development of broadly neutralizing antibodies to HIV.[140]

☼ Inference of point mutation models is just starting to use methods with a probabilistic foundation, and more work needs to be done. One outstanding challenge is that the process of somatic hypermutation happens on phylogenetic trees, and it is difficult to do model inference on phylogenetic trees with context-sensitive models. Indeed, phylogenetic model inference typically integrates out potential internal states as part of the model fitting process on the tree; this is enabled by the use of the Felsenstein algorithm which requires an independence-among-sites assumption (more details below). That assumption is of course violated for context-sensitive models. Our group has used an additional sampling step to marginalize out the possible ancestral sequences of a given sequence, and avoid the need to do so in a fully phylogenetic context by selecting only one sequence per clonal family (ie, phylogenetic tree). Such estimation has been previously done for simpler classes of models.[141,142]

We are not aware of any probabilistic models for indels specifically in the somatic hypermutation process. For molecular sequences in general, such models first appeared in 1986,[143] followed by the foundational TKF models.[144,145] Recent work has defined a class of indel models with attractive computational properties.[146,147]

☾ A more biologically explicit mutation model would consider AID deamination and repair processes in terms of mismatch repair, base excision repair, indel introduction, and gene conversion.[116] Although a fully specified mechanistic model would be challenging for efficient inference, it is certainly suitable for simulation. Our group is currently using our more flexible mutation modeling setup[24] to "fish out" certain types of effects that will allow us to estimate rates of these various pathways, and form the foundation for such models.

2.7 | B-cell clonal expansion: lineage development

B cells undergo a Darwinian process of mutation and selection in the germinal center to improve binding to antigen. Each B cell entering the germinal center founds a lineage (realized as a phylogenetic tree), and is the unmutated ancestor of all of its mutated descendants. Mutation and selection happen in the dark and light zones of germinal centers, respectively: in the dark zone B cells reproduce, introducing additional diversity by the somatic hypermutation process described above, whereas in the light zone B cells compete to retrieve antigen from follicular dendritic cells. Recent work has emphasized the importance of T cell help from retrieving antigen as opposed to direct stimulation to reproduce from BCR crosslinking.[31]

Affinity maturation is a dynamic population-level process. Using a mouse engineered to express a reporter of apoptosis, researchers have found that apoptosis is the "default" outcome in the absence of T-cell help.[148] Intensive examination of individual germinal centers has led to the hypothesis of "clonal bursts" in which B cells divide in rapid succession due to a strong T-cell stimulus.[149] Despite what would seem to be a very strong selective environment, phylogenetic analysis combined with affinity measurements has not revealed a steady march toward increased affinity in sampled germinal centers.[149,150] Existing antibodies and B cells, including those appearing during the germinal center reaction[151] and those from prior exposures,[152] change the evolutionary dynamics of the germinal center reaction.

Germinal centers are not seeded by single naive cells. Indeed, random florescence labeling shows that many cells initially seed germinal centers, although these germinal centers often "resolve" to the descendants of a single cell through competition.[149] In addition, B cells entering the germinal center need not be naive: mathematical simulation[153] and mutation analysis of vaccination studies in mice[154] support the hypothesis that lineages can be re-seeded from existing lineages.

So much previous work has been done analyzing B-cell sequence lineage development, that we will divide this section into further mini-sections.

2.8 | Clonal family inference

Many computational methods have been developed to reconstruct the hidden aspects of B-cell clonal expansion and infer the dynamics behind it. Any bulk sample of B cells mixes sequences deriving from different naive cells and responding to different antigens. Thus, an important first step for analysis is to group sequences into "clonal families," namely, collections of sequences that descended from a single naive cell. The most popular means of doing this is to apply single-linkage clustering to the sequences, allowing sequences to cluster if they are annotated to have the same V and J sequences, have the same CDR3 length, and are less than some fixed Hamming distance apart. Needless to say, there are issues with each of these assumptions. Somatic hypermutation may cause uncertainty as to germline gene assignment, and insertion/deletion mutations (indels) may change CDR3 length. However, the assumption of a fixed cutoff, even a per-repertoire fixed cutoff,[155] seems the most problematic. The most obvious counter-example to this assumption is given by broadly neutralizing antibodies against HIV, which with around 100 mutations have the same order of divergence from germline genes as these germline genes have to one another.[156] From a phylogenetic perspective, fixed-cutoff methods make the surprising assumption that branch lengths in the process of somatic hypermutation cannot be longer than some fixed quantity. This assumption seems even less sensible when we consider that repertoires are small samples from a large population; when we drop leaves from a phylogenetic tree because of sampling, the resulting branches become longer.

To avoid such assumptions, our group has developed a likelihood-based means of inferring clonal families in our partis software package.[7] We begin by recasting the problem to one of inferring groups of sequences that have the same naive sequence. This differs from the original question of inferring clonal families, because the same naive sequence can be generated by two different rearrangement events. To solve this question, ideally one would do a perfect job of inferring a naive sequence from each mature sequence and then simply cluster based on those inferred naive sequences. However, such a procedure is not possible because there are many ways to obtain a given sequence from different ancestors via somatic hypermutation. For this reason, the method calculates a likelihood that two groups of sequences come from the same naive ancestor, while integrating over possible naive sequences. By comparing this likelihood to the alternative hypothesis that the two groups do not share ancestry via a likelihood ratio, one is able to decide whether these two groups should be merged into one. The method applies this likelihood-based framework via agglomerative clustering in a manner reminiscent of the neighbor-joining algorithm.[157,158] A naive implementation of this procedure would be far too slow for actual use, and thus, the method uses many optimizations. Some of these come without any drop in accuracy, whereas some strike a balance between computational tractability and accuracy.

Inferences of such clusters will always be an uncertain process, which invites a Bayesian approach to obtain posterior distributions on the clusters. Indeed, early unpublished versions of our

procedure did this hierarchical agglomeration via sequential Monte Carlo (SMC),[159] an algorithm that can be thought of as a probabilistically correct type of genetic algorithm. In SMC one maintains a population of objects being inferred, and at each stage makes some modification. In this case, our software maintained a population of different partial clusterings, and at each stage every partial clustering makes some probabilistic merge weighted by a likelihood ratio.

This procedure was too slow and cumbersome to be applied to large sequence datasets. However, our group experimented with it enough to feel confident that uncertainty was basically "one-dimensional," such that the primary unknown quantity was the degree of clustering. Given that partis records the sequence of clusterings that lead to each inferred cluster along with their likelihoods, we left our Bayesian ambition there. A Bayesian clustering algorithm based on a Dirichlet process mixture model has been described[160] although this algorithm does not appear to have been applied in practice.

2.9 | Phylogenetic inference on B-cell sequences

Even once the clusters are fixed, estimating the tree for each cluster is non-trivial. Besides the fact that estimating a phylogenetic tree is an inherently hard problem, B-cell sequences have features that differentiate them from typical applications of phylogenetics, and thus require special algorithms. When sampling is dense, it is not unusual to sample ancestor-descendant pairs. (Even if we are not actually sampling the true ancestor of a given cell we may sequence a cell that is identical to it.) The relatively short branch lengths between sequences has motivated an extensive use of parsimony,[161,162] a method in which one chooses the tree that minimizes the number of mutations required to explain observed sequence data at the tips. It is important to restrict the use of parsimony to cases with short branch lengths as it is known to be statistically inconsistent when branches become long[163]; that is, it will produce the incorrect tree with probability one in the limit of long sequences.

When single-cell sequencing is applied to a densely sequenced sample, such as one from a germinal center,[149] each sequence comes equipped with a meaningful abundance. This information can be productively used to guide phylogenetic inference.[164] The intuition behind this approach is that, first, sampled abundance reflects the overall abundance of that genotype in the population, and second, more frequent cells are more likely to leave mutant descendants. For this reason, we should prefer trees that connect descendants to more frequently observed ancestors over those that do not.

Substantial information about the ancestral sequence can be inferred with knowledge of germline sequences. Indeed, if one knows that a given V gene was used in the process of VDJ recombination, we know the ancestral state for that region of the sequence. This contrasts most applications of phylogenetics, in which ancestral states are typically unknown. In order to integrate this information one needs a computational framework that knows about both VDJ rearrangement and phylogenetics. Kepler has described such an approach, which iteratively infers a tree while estimating a posterior on unmutated ancestor sequences.[165] Each iteration takes the unmutated ancestral sequence with the highest posterior probability, builds a tree using that sequence at the root, and then re-estimates the posterior on the unmutated ancestor.

The highly context-sensitive mutation processes found in somatic hypermutation (reviewed above) violate the near-universal phylogenetic assumption of independent evolution between sites. This assumption is essential for efficient likelihood computation in phylogenetics via the Felsenstein algorithm.[17] This can be understood intuitively as follows: if the substitution history at the first site depends on the second, and the history at the second depends on the third, then continuing this string of dependencies means that we must consider evolution to be happening on the whole sequence at a time. This is computationally intractable as the state space for nucleotides is four to the power of the sequence length, defeating the traditional use of transition matrices.

Thus if one wants to stay inside the usual likelihood-based framework for phylogenetics one must use approximations to maintain the independence assumption. An important step forward was recently made by incorporating context information into a codon model.[166] In codon models, one considers codons, rather than individual nucleotides, to be the units of evolution and assumes independence between those codons. By averaging out the part of nucleotide contexts that extend beyond the codon boundary, this work maintains a model that has independence between codons. This approach has the additional advantage that one can estimate parameters of selection and context sensitivity directly from the model.

B-cell sequence analysis has more emphasis on phylogenetic ancestral sequence inference than is typical for other applications of phylogenetics, and for good reason. Ancestral sequence inference methods enable a beautiful convergence of computational analysis and laboratory experiments: estimated ancestral sequences can be expressed and built in the lab to test their properties.[156,167,168] Such experiments, when combined with structural analysis, give real insight into how substitutions lead to improved affinity. The computational tool for these analyses has typically been PHYLIP,[169] although other programs[170,171] are faster or have additional features.

2.10 | Selection inference

We can get additional insight into the evolutionary process by estimating the strength of natural selection on a collection of sequences using codon-based methods. Such methods make inferences by considering the relative rate of synonymous (between codons for an amino acid) to non-synonymous (between amino acid) substitution. The intuition is that if there is selection to preserve an amino acid, one will see an excess of synonymous changes compared to non-synonymous ones because non-synonymous changes will be selected out of the population. The opposite will hold for cases when amino acid change is beneficial.

Such analysis is made difficult by the context-sensitive mutation process: because the probability of substitution is influenced by the local sequence context on one hand, and natural selection on codons

on the other, false conclusions can be drawn if one does not correct for it explicitly.[126] Such correction is indeed possible.[127,128,172–174] Repertoire-level selection has been measured in the CDR region versus the framework region[127,172,173] and in the "trunk" (edges leading from the naive ancestor to the most recent common ancestor of sampled sequences) versus the rest of the tree,[174] with results broadly consistent between individuals. This theme of consistency is even more striking on a per-codon level,[128] which shows diverse amounts of selection at various sites in the framework region that are consistent among individuals.

Tree shape and structure have also been used to estimate selective pressure on ensembles of trees. Early work used graph-theoretic properties of trees to estimate selection strength[175]; correlation between these measures and selection strength was determined by simulation.[29] Later authors found that such properties can be distorted by difficult-to-control experimental factors.[176] They proposed an alternative method mapping mutations onto the edges of the tree and using patterns of replacement and silent nucleotide substitutions filtered to only include substitutions on non-terminal branches.[176]

A more ambitious goal is to estimate selection on a single tree at a time.

One recent approach compares tree balance (the number of descendants on one side of a node vs another) at nodes directly below edges with amino acid changes vs those without.[177] An investigation of vaccine-responsive trees[178] applied local branching rates[179] and a more classical investigation of site-frequency spectra[180] to look for evidence of selective sweeps.

2.11 | Modeling lineage development

The dynamic evolution of antibodies in germinal centers has been modeled for over a quarter century, for example leading to an early prediction of re-entry of circulating B cells back into the germinal center.[30] An early computer simulation framework, "Clone," although not explicitly simulating an actual molecular sequence, simulated patterns of mutation in various parts of the BCR and their consequences.[181] Others have performed ABC-like (see first section for an introduction to ABC) analyses where they fit values such as mutation rate, selection, and clone affinity based on concordance of summary statistics.[27–29] These analyses have typically been independent of existing population genetics theory, although recent work[178] makes use of site-frequency spectrum tools from population genetics. Another vein of work uses agent-based and differential equation-based modeling to iteratively improve compartmental models of B-cell development.[182–191] For chronic infections such as HIV, antibody-pathogen coevolution certainly plays a role[192] although the dynamics between antibody emergence and viral escape are difficult to pin down.[193] Some researchers have found a "trunk-canopy" tree structure from mature sequence data, in which a long "trunk" branch from the root extends from the naive sequence, after which there is a "canopy" of diversification.[174] However, it has been pointed out that the extent to which this structure is seen depends on the level of clustering.[194]

☼ The previous review shows the disjointed state of the field: although clonal clustering, phylogenetics, selection inference, and modeling are all describing aspects of the same underlying process, they are divided into different problems (note that rearrangement inference, which is closely tied in with phylogenetic estimation, was relegated to its own section above, whereas isotype, which is closely tied with mutation processes on trees, appears in the next section!). We must work toward unifying these various aspects in a shared framework.

Bayesian statistics offers a coherent framework for such information sharing and integration over uncertain latent states. Although estimation of these complex posteriors will not be easy, we will be rewarded by more accurate inferences, leading to a more complete understanding of how affinity maturation works. Our group is currently building on prior work[165] to develop a Bayesian sampling procedure on trees that integrates out uncertainty in the unmutated common ancestor using a hidden Markov model.

We are also inspired by the work of Jonathan Laserson et al[160,195] who describes how sampling ancestral sequences explicitly as part of an MCMC can actually increase efficiency. This echoes earlier work in a more general setting.[196] The value of such sampling will be even greater when using more complex context-sensitive models, for which calculating likelihoods currently requires more intensive extensions to Gibbs sampling procedures (eg, that of Chib[197]) to compute the marginal likelihood. Other types of analysis, such as that of selection pressure,[128] also require ancestral sequence inference. Thus we believe that the next generation of phylogenetic algorithms for BCR sequences will infer a joint posterior on ancestral sequences and trees.

We also believe that tree-valued stochastic models will provide a unified foundation for learning about the diversification process from B-cell sequence data. Researchers working on viral populations have developed sophisticated tools for learning about viral spread by estimating ancestral population size using Bayesian "skyline" analysis[198] and phylogenetic generalized linear models.[15] Somewhat analogous stochastic models for B-cell development have been devised, but have only been used to generate distributions of summary statistics rather than being used for inference.[28] A more powerful tactic will be to develop models with parameters of interest and perform parameter inference directly.

Although repertoire-scale inference with a "dream" algorithm getting posterior distributions of all relevant parameters will not be possible, we can scale our computational ambition to the question at hand. If we are very interested in a specific clonal family, it may be worth expending considerable computational effort in order to get high quality inferences for that family. This may include probabilistically sampling alternative clusterings of that clonal family. On the other hand, if we are looking for repertoire-level characteristics we will want to scale back our effort on each individual family in order to get an overall picture (although it is important that such algorithms are unbiased).

☾ Realistic forward-time models are essential to help guide the design and implementation of inferential algorithms. For example, there is currently a need for models of affinity maturation that

generate nucleotide sequences and trees interdependently with some level of realism. Although antibody affinity models are relatively old[199] and plentiful (see above), we are not aware of any that generate nucleotide sequences. Our group is currently developing such a model as part of a benchmarking exercise of ancestral reconstruction methods.

2.12 | B-cell clonal expansion: Isotype

Antibodies have an isotype-determining constant region that establishes the function of the antibody in the immune system. Isotype can change through class-switch recombination, which arises due to double-stranded breaks resulting from AID deamination.[116,200] High-throughput sequencing including isotype information is now available, and is shining light on this process. For example, there are significant differences between isotypes in terms of their levels of somatic hypermutation.[201] These new data are also elucidating the rate with which antibodies switch isotype classes.[202,203] An analysis of sister lineages on either side of a branch point has suggested that the probability of switching to the various other isotypes is determined by more than just the current isotype[202]: rather, there is some additional hidden factor that determines the switching probability. This could be summarized by saying that the isotype-switching process does not satisfy the Markov property.

☼ If we do assume the Markov property, one can formulate an isotype model using existing continuous-time trait models. Inference under such models is well developed from both maximum likelihood[204] and Bayesian[205] perspectives. Adding isotype as a hidden state in phylogenetic inference would be straightforward.

☾ One may also wish to model a non-Markov latent state for which existing inferential techniques will not apply. Another interesting type of model would e one in which mutation and isotype-switching are linked probabilistically.

2.13 | Estimating the complete adaptive immune response

Although repertoire sequencing offers a remarkable perspective into the complex process of immune state, it will always offer an incomplete picture of the immune response. First, it is well-acknowledged that we are taking a small sample from a very large population. As such, it is common to extrapolate the total number of unique immune receptors from a sample.[206]

However, this is not the whole story: the common practice of sequencing from blood may not reflect what is happening with B and T cells in other compartments. Recent work is beginning to lay the foundation for understanding the whole B- and T-cell response from blood samples. This has included a "B-cell atlas" of samples from many tissues of organ donors,[207] as well as sequential fine needle aspirates from rhesus,[208] and sequencing from individual germinal centers using lymph node dissection in mice.[149]

In another direction one would like to understand the essential role that circulating antibodies play in the immune response. Although B-cell sequencing gives some idea of what antibodies can be made, it is certainly not the same as assaying the antibodies present in an individual. The soluble antibody repertoire is determined by expression and antibody lifetime. To do this, recent work has combined protein mass spectrometry with antibody sequencing.[209,210] Hopefully new protein sequencing methods[211] will expand our perspective on soluble antibodies.

One may continue along these lines and say that even the pool of circulating antibodies are not the most interesting factor, and rather one should be interested in the collection of antigens that can be bound by those antibodies. For this, recent work has used antigen microarrays[212,213] to infer what peptides can be bound by circulating antibodies. For T cells, yeast display has been used to identify the peptide specificity of TCRs found in cancer.[214] Abstracting one notch further, one can use immunological assays between viral strains to assay an antigenic "distance" between them[215,216] that captures cross-reactivity of antibodies.

These complexities are well known to theoreticians. The doctrine of "original antigenic sin" is over 60 years old[217] and modern methods continue to support past exposures as being essential for future development.[152,218,219] Perhaps the closest analysis of actual sequences are models of population-level immunity in which the fitness of a given influenza sequence is in part determined by its similarity to existing sequences to which the population is already presumably immune.[220,221] There are also controlled experiments and mathematical models working to understand the impact of antibody feedback,[151,188,189,222,223] although this work has not been generalized to an inferential framework that can be used to understand individual repertoire datasets. "Mutational antigenic profiling,"[224–226] which reveals how mutating an antigen can change antibody binding, and "deep mutational scanning," used to understand the impact of antibody sequence variation on binding,[227,228] may be helpful in these efforts.

☼ Given a lot of "B-cell atlas" type data, one might be able to develop a migration model between the various compartments and infer rates based on observations of the same or related clones in different compartments. The challenge with such a project will be to untangle re-seeding from early seeding and partial persistence. Also, using such data, one may be able to model cell population sizes of difficult-to-sample compartments from ones that are easier to sample.

☾ One could dream of a model that attempts to capture the antigenic space that is covered by existing circulating antibodies. In particular, recent efforts introduce antibody landscapes in the context of influenza.[229]

3 | CONCLUSION

We have reviewed opportunities for probabilistic modeling in B- and T-cell sequence analysis. To summarize, probabilistic models have a lot to contribute to rearrangement and lineage inference. However, inferences on the functional properties of individual sequences (such as for initial selective filters or antigen binding) seem better

done with machine learning methods rather than generative models for which likelihood calculation is not tractable. For those aspects well suited to probabilistic modeling, we will be rewarded for integrating various aspects into a single framework where one level can communicate important information, including uncertainty, to another level. For example, from B-cell sequence analysis:

- There is considerable signal in patterns of shared mutation that can help guide clustering inference, and the correct way of doing so is to combine phylogenetic inference with clustering.
- It is common to include an inferred, unmutated common ancestor into a sequence alignment for phylogenetic inference. Accounting for the corresponding uncertainty is important to gain accurate inferences on the processes that led to the observed sequences.
- Phylogenetic trees are also uncertain, and disregarding that uncertainty will skew our downstream analyses of selection and models.

This model hierarchy can extend beyond the single-sample level to individual-level analysis through time, or population-level analysis. The parameters we learn from these larger studies, such as germline gene existence and frequency, can feed back down to improve per-sample analysis. They can also be used to analyze predictors of individual-level immune variation.[230]

The computational statistician interested in immune receptor modeling is blessed with a complex biological system to analyze, intractable computational problems heaped on top of one another, and an ever-expanding collection of datasets generated from various in vivo and in vitro perturbations. New methods are needed to perform inference under complex hierarchical models of immune receptor development for the optimistic program laid out in this paper to become a reality. Although the field of computational immunology dates back many decades, we can gain inspiration and adapt techniques from the even longer tradition of macroevolutionary and ecological theory. There, we have seen a complex interplay of generative models, summary statistics, and inferential models that have enabled the field's progress.

ACKNOWLEDGEMENTS

We thank the AIRR community (http://airr-community.org) for discussions that have greatly enriched our understanding and appreciation of the field. Christian Busse, Kristian Davidsen, William DeWitt, Julia Fukuyama, David Shaw, Duncan Ralph, and Corey Watson provided helpful feedback on this manuscript. We thank Curt Callan for allowing us to use parts of his figure in our Figure 2.

CONFLICT OF INTEREST

We declare no conflict of interest.

REFERENCES

1. Sompayrac LM. *How the immune system works*. Hoboken, NJ: John Wiley & Sons; 2011.
2. Robins H. Immunosequencing: Applications of immune repertoire deep sequencing. *Curr Opin Immunol*. 2013;25:646-652.
3. Yaari G, Kleinstein SH. Practical guidelines for B-cell receptor repertoire sequencing analysis. *Genome Med*. 2015;7:121.
4. Murugan A, Mora T, Walczak AM, Callan CG Jr. Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proc Natl Acad Sci USA*. 2012;109:16161-16166.
5. Elhanati Y, Sethna Z, Marcou Q, Callan CG Jr, Mora T, Walczak AM. Inferring processes underlying B-cell repertoire diversity. *Philos Trans R Soc Lond B Biol Sci*. 2015;370:20140243.
6. Ralph DK, Matsen FA IV. Consistency of VDJ Rearrangement and Substitution Parameters Enables Accurate B Cell Receptor Sequence Annotation. *PLoS Comput Biol*. 2016;12:e1004409.
7. Ralph DK, Matsen FA IV. Likelihood-based inference of B cell clonal families. *PLoS Comput Biol*. 2016;12:e1005086.
8. MacKay DJ. *Information Theory, Inference, and Learning Algorithms*. Cambridge, UK: Cambridge University Press; 2003.
9. Hornik K, Leisch F, Zeileis A. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of DSC (Vol 2)*. r-project.org; 2003:1-1.
10. Lunn D, Spiegelhalter D, Thomas A, Best N. The BUGS project: Evolution, critique and future directions. *Stat Med*. 2009;28:3049-3067.
11. Carpenter B, Gelman A, Hoffman M, et al. Stan: A probabilistic programming language. *J Stat Softw*. 2016;20:1-37.
12. Nascimento FF, dos Reis M, Yang Z. A biologists guide to Bayesian phylogenetic analysis. *Nature Ecology & Evolution*. 2017;1:1446.
13. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol*. 2012;29:1969-1973.
14. Ronquist F, Teslenko M, van der Mark P, et al. MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol*. 2012;61:539-542.
15. Dudas G, Carvalho LM, Bedford T, et al. Virus genomes reveal factors that spread and sustained the Ebola epidemic. *Nature*. 2017;544:309-315.
16. Efron B. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge, UK: Cambridge University Press; 2012.
17. Felsenstein J. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J Mol Evol*. 1981;17:368-376.
18. Roch S. A short proof that phylogenetic tree reconstruction by maximum likelihood is hard. *IEEE/ACM Trans Comput Biol Bioinform*. 2006;3:92-94.
19. Durbin R, Eddy SR, Krogh A, Mitchison G. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge, UK: Cambridge University Press; 1998.
20. Scott SL. Bayesian methods for hidden Markov models. *J Am Stat Assoc*. 2011;457:337-351.
21. Rogozin IB, Kolchanov NA. Somatic hypermutagenesis in immunoglobulin genes. II. Influence of neighbouring base sequences on mutagenesis. *Biochim Biophys Acta*. 1992;1171:11-18.
22. Yaari G, Vander Heiden JA, Uduman M, et al. Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput immunoglobulin sequencing data. *Front Immunol*. 2013;4:358.
23. Wei GCG, Tanner MA. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J Am Stat Assoc*. 1990;85:699-704.
24. Feng J, Shaw DA, Minin VN, Simon N, Matsen FA IV. *Survival Analysis of DNA Mutation Motifs With Penalized Proportional Hazards. arXiv*. 2017;arXiv:1711.04057.
25. Marin JM, Pudlo P, Robert CP, Ryder RJ. Approximate Bayesian computational methods. *Stat Comput*. 2012;22:1167-1180.

26. Robert PA, Rastogi A, Binder SC, Meyer-Hermann M. How to simulate a germinal center. In: Calado DP, ed. *Germinal Centers: Methods and Protocols*. *Methods in Molecular Biology*. New York, NY: Springer New York; 2017:303-334.
27. Kleinstein SH, Louzoun Y, Shlomchik MJ. Estimating hypermutation rates from clonal tree data. *J Immunol*. 2003;171:4639-4649.
28. Magori-Cohen R, Louzoun Y, Kleinstein SH. Mutation parameters from DNA sequence data using graph theoretic measures on lineage trees. *Bioinformatics*. 2006;22:e332-e340.
29. Shahaf G, Barak M, Zuckerman NS, Swerdlin N, Gorfine M, Mehr R. Antigen-driven selection in germinal centers as reflected by the shape characteristics of immunoglobulin gene lineage trees: A large-scale simulation study. *J Theor Biol*. 2008;255:210-222.
30. Kepler TB, Perelson AS. Cyclic re-entry of germinal center B cells and the efficiency of affinity maturation. *Immunol Today*. 1993;14:412-415.
31. Victora GD, Schwickert TA, Fooksman DR, et al. Germinal center dynamics revealed by multiphoton microscopy with a photoactivatable fluorescent reporter. *Cell*. 2010;143:592-605.
32. Mesin L, Ersching J, Victora GD. Germinal center B cell dynamics. *Immunity*. 2016;45:471-482.
33. Blei DM. Build, compute, critique, repeat: Data analysis with latent variable models. *Annu Rev Stat Appl*. 2014;1:203-232.
34. Bassing CH, Swat W, Alt FW. The mechanism and regulation of chromosomal V(D)J recombination. *Cell*. 2002;109(Suppl):S45-S55.
35. Watson CT, Breden F. The immunoglobulin heavy chain locus: Genetic variation, missing data, and implications for human disease. *Genes Immun*. 2012;13:363-373.
36. Jackson KJL, Kidd MJ, Wang Y, Collins AM. The shape of the lymphocyte receptor repertoire: Lessons from the B cell receptor. *Front Immunol*. 2013;4:263.
37. Rowen L, Koop BF, Hood L. The complete 685-kilobase DNA sequence of the human beta T cell receptor locus. *Science*. 1996;272:1755-1762.
38. Boysen C, Simon MI, Hood L. Analysis of the 1.1-Mb human α/δ T-cell receptor locus with bacterial artificial chromosome clones. *Genomecshlporg*. 1997;7:330-338.
39. Watson CT, Steinberg KM, Huddleston J, et al. Complete haplotype sequence of the human immunoglobulin heavy-chain variable, diversity, and joining genes and characterization of allelic and copy-number variation. *Am J Hum Genet*. 2013;92:530-546.
40. Watson CT, Steinberg KM, Graves TA, et al. Sequencing of the human IG light chain loci from a hydatidiform mole BAC library reveals locus-specific signatures of genetic diversity. *Genes Immun*. 2014;16:24-34.
41. Feeney AJ, Atkinson MJ, Cowan MJ, Escuro G, Lugo G. A defective Vkappa A2 allele in Navajos which may play a role in increased susceptibility to haemophilus influenzae type B disease. *J Clin Invest*. 1996;97:2277-2282.
42. Wang Y, Jackson KJ, Gäeta B, et al. Genomic screening by 454 pyrosequencing identifies a new human IGHV gene and sixteen other new IGHV allelic variants. *Immunogenetics*. 2011;63:259-265.
43. Gadala-Maria D, Yaari G, Uduman M, Kleinstein SH. Automated analysis of high-throughput B-cell sequencing data reveals a high frequency of novel immunoglobulin V gene segment alleles. *Proceedings of the National Academy of Sciences of the USA*. 2015;112:E862-E870.
44. Scheepers C, Shrestha RK, Lambson BE, et al. Ability to develop broadly neutralizing HIV-1 antibodies is not restricted by the germline Ig gene repertoire. *J Immunol*. 2015;194:4371-4378.
45. Watson CT, Glanville J, Marasco WA. The individual and population genetics of antibody immunity. *Trends Immunol*. 2017;38:459-470.
46. Avnir Y, Watson CT, Glanville J, et al. IGHV1-69 polymorphism modulates anti-influenza antibody repertoires, correlates with IGHV utilization shifts and varies by ethnicity. *Sci Rep*. 2016;6:20842.
47. Boyd SD, Gaëta BA, Jackson KJ, et al. Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. *J Immunol*. 2010;184:6986-6992.
48. Kidd MJ, Chen Z, Wang Y, et al. The inference of phased haplotypes for the immunoglobulin H chain V region gene loci by analysis of VDJ gene rearrangements. *J Immunol*. 2012;188:1333-1340.
49. Corcoran MM, Phad GE, Vázquez Bernat N, et al. Production of individualized V gene databases reveals high levels of immunoglobulin genetic diversity. *Nat Commun*. 2016;7:13642.
50. Ralph DK, Matsen FA IV. *Per-Sample Immunoglobulin Germline Inference From B Cell Receptor Deep Sequencing Data*. *arXiv*. 2017; arXiv:1711.05843.
51. Kirik U, Greiff L, Levander F, Ohlin M. Parallel antibody germline gene and haplotype analyses support the validity of immunoglobulin germline gene inference and discovery. *Mol Immunol*. 2017;87:12-22.
52. Reams AB, Roth JR. Mechanisms of gene duplication and amplification. *Cold Spring Harb Perspect Biol*. 2015;7:a016592.
53. Luo S, Yu JA, Li H, Song YS. Worldwide genetic variation of the IGHV and TRBV immune receptor gene families in humans. *BioRxiv*. 2017. https://doi.org/10.1101/155440
54. Luo S, Yu JA, Song YS. Estimating copy number and allelic variation at the immunoglobulin heavy chain locus using short reads. *PLoS Comput Biol*. 2016;12:e1005117.
55. Schatz DG, Ji Y. Recombination centres and the orchestration of V (D) J recombination. *Nat Rev Immunol*. 2011;11:251-263.
56. Cowell LG, Davila M, Yang K, Kepler TB, Kelsoe G. Prospective estimation of recombination signal efficiency and identification of functional cryptic signals in the genome by statistical modeling. *J Exp Med*. 2003;197:207-220.
57. Montefiori L, Wuerffel R, Roqueiro D, et al. Extremely long-range chromatin loops link topological domains to facilitate a diverse antibody repertoire. *Cell Rep*. 2016;14:896-906.
58. Kepler TB, Borrero M, Rugerio B, McCray SK, Clarke SH. Interdependence of N nucleotide addition and recombination site choice in V(D)J rearrangement. *J Immunol*. 1996;157:4451-4457.
59. Volpe JM, Kepler TB. Large-scale analysis of human heavy chain V(D)J recombination patterns. *Immunome Res*. 2008;4:3.
60. Elhanati Y, Marcou Q, Mora T, Walczak AM. repgenHMM: A dynamic programming tool to infer the rules of immune receptor generation from sequence data. *Bioinformatics*. 2016;32:1943-1951.
61. Sethna Z, Elhanati Y, Dudgeon CS, et al. Insights into immune system development and function from mouse T-cell repertoires. *Proc Natl Acad Sci USA*. 2017;114:2253-2258.
62. Wilson PC, Wilson K, Liu YJ, Banchereau J, Pascual V, Capra JD. Receptor revision of immunoglobulin heavy chain variable region genes in normal human B lymphocytes. *J Exp Med*. 2000;191:1881-1894.
63. Collins AM, Ikutani M, Puiu D, Buck GA, Nadkarni A, Gaeta B. Partitioning of rearranged Ig genes by mutation analysis demonstrates D-D fusion and V gene replacement in the expressed human repertoire. *J Immunol*. 2004;172:340-348.
64. Meng W, Jayaraman S, Zhang B, et al. Trials and tribulations with VH replacement. *Front Immunol*. 2014;5:10.
65. Ohm-Laursen L, Nielsen M, Larsen SR, Barington T. No evidence for the use of DIR, D-D fusions, chromosome 15 open reading frames or VHreplacement in the peripheral repertoire was found on application of an improved algorithm, JointML, to 6329 human immunoglobulin H rearrangements. *Immunology*. 2006;119:265-277.
66. Lee DW, Khavrutskii I, Wallqvist A, Bavari S, Cooper CL, Chaudhury S. BRILIA: Integrated tool for high-throughput annotation and lineage tree assembly of B-cell repertoires. *Front Immunol*. 2016;7:681.
67. Gaëta BA, Malming HR, Jackson KJL, Bain ME, Wilson P, Collins AM. iHMMune-align: Hidden Markov model-based alignment and

identification of germline genes in rearranged immunoglobulin gene sequences. *Bioinformatics*. 2007;23:1580-1587.
68. Munshaw S, Kepler TB. SoDA2: A Hidden Markov Model approach for identification of immunoglobulin rearrangements. *Bioinformatics*. 2010;26:867-872.
69. Marcou Q, Mora T, Walczak AM. High-throughput immune repertoire analysis with IGoR. *Nat Commun*. 2018;9:561.
70. Kidd MJ, Jackson KJL, Boyd SD, Collins AM. DJ pairing during VDJ recombination shows positional biases that vary among individuals with differing IGHD locus immunogenotypes. *J Immunol*. 2015;196:1158-1164.
71. Nadel B, Feeney AJ. Influence of coding-end sequence on coding-end processing in V(D)J recombination. *J Immunol*. 1995;155:4322-4329.
72. Nadel B, Feeney AJ. Nucleotide deletion and P addition in V(D)J recombination: A determinant role of the coding-end sequence. *Mol Cell Biol*. 1997;17:3768-3778.
73. Larimore K, McCormick MW, Robins HS, Greenberg PD. Shaping of human germline IgH repertoires revealed by deep sequencing. *J Immunol*. 2012;189:3221-3230.
74. Meng W, Yunk L, Wang LS, et al. Selection of individual VH genes occurs at the pro-B to pre-B cell transition. *J Immunol*. 2011;187:1835-1844.
75. Benichou J, Glanville J, Prak ETL, et al. The restricted DH gene reading frame usage in the expressed human antibody repertoire is selected based upon its amino acid content. *J Immunol*. 2013;190:5567-5577.
76. Elhanati Y, Murugan A, Callan CG Jr, Mora T, Walczak AM. Quantifying selection in immune receptor repertoires. *Proc Natl Acad Sci USA*. 2014;111:9875-9880.
77. DeKosky BJ, Ippolito GC, Deschner RP, et al. High-throughput sequencing of the paired human immunoglobulin heavy and light chain repertoire. *Nat Biotechnol*. 2013;31:166-169.
78. Howie B, Sherwood AM, Berkebile AD, et al. High-throughput pairing of T cell receptor α and β sequences. *Sci Transl Med*. 2015;7:301ra131.
79. Grigaityte K, Carter JA, Goldfless SJ, et al. *Single-Cell Sequencing Reveals αβ Chain Pairing Shapes the T Cell Repertoire*. *BioRxiv*. 2017. https://doi.org/10.1101/213462
80. Emerson RO, DeWitt WS, Vignali M, et al. Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nat Genet*. 2017;49:659-665.
81. Desponds J, Mora T, Walczak AM. Fluctuating fitness shapes the clone-size distribution of immune repertoires. *Proc Natl Acad Sci USA*. 2016;113:274-279.
82. Desponds J, Mayer A, Mora T, Walczak AM. Population dynamics of immune repertoires. *BioRxiv*. 2017. https://doi.org/10.1101/112755
83. Qi Q, Cavanagh MM, Le Saux S, et al. Diversification of the antigen-specific T cell receptor repertoire after varicella zoster vaccination. *Sci Transl Med*. 2016;8:332ra46.
84. Pogorelyy MV, Elhanati Y, Marcou Q, et al. Persisting fetal clonotypes influence the structure and overlap of adult human T cell receptor repertoires. *PLoS Comput Biol*. 2017;13:e1005572.
85. Boyd SD, Liu Y, Wang C, Martin V, Dunn-Walters DK. Human lymphocyte repertoires in ageing. *Curr Opin Immunol*. 2013;25:511-515.
86. Dash P, Fiore-Gartland AJ, Hertz T, et al. Quantifiable predictive features define epitopespecific T cell receptor repertoires. *Nature*. 2017;547:89-93.
87. Yokota R, Kaminaga Y, Kobayashi TJ. Quantification of intersample differences in T-cell receptor repertoires using sequence-based information. *Front Immunol*. 2017;8:1500.
88. Glanville J, Huang H, Nau A, et al. Identifying specificity groups in the T cell receptor repertoire. *Nature*. 2017;547:94-98.
89. Sharon E, Sibener LV, Battle A, Fraser HB, Garcia KC, Pritchard JK. Genetic variation in MHC proteins is associated with T cell receptor expression biases. *Nat Genet*. 2016;48:995-1002.
90. Thomas N, Best K, Cinelli M, et al. 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*. 2014;30:3181-3188.
91. Ostmeyer J, Christley S, Rounds WH, et al. Statistical classifiers for diagnosing disease from immune repertoires: A case study using multiple sclerosis. *BMC Bioinformatics*. 2017;18:401.
92. Pogorelyy MV, Minervina AA, Chudakov DM, et al. Method for identification of condition-associated public antigen receptor sequences. *Elife*. 2018;7:e33050.
93. Shugay M, Bagaev DV, Zvyagin IV, et al. VDJdb: A curated database of T-cell receptor sequences with known antigen specificity. *Nucleic Acids Res*. 2017;46:D419-D427.
94. Birnbaum ME, Mendoza JL, Sethi DK, et al. Deconstructing the peptide-MHC specificity of T cell recognition. *Cell*. 2014;157:1073-1087.
95. Victora GD, Nussenzweig MC. Germinal centers. *Annu Rev Immunol*. 2012;30:429-457.
96. Manz RA, Hauser AE, Hiepe F, Radbruch A. Maintenance of serum antibody levels. *Annu Rev Immunol*. 2005;23:367-386.
97. Galson JD, Kelly DF, Truck J. Identification of antigen-specific B-cell receptor sequences from the total B-cell repertoire. *Crit Rev Immunol*. 2015;35:463-478.
98. Jackson KJL, Liu Y, Roskin KM, et al. Human responses to influenza vaccination show seroconversion signatures and convergent antibody rearrangements. *Cell Host Microbe*. 2014;16:105-114.
99. Laserson U, Vigneault F, Gadala-Maria D, et al. High-resolution antibody dynamics of vaccine-induced immune responses. *Proc Natl Acad Sci USA*. 2014;111:4928-4933.
100. Jiang N, He J, Weinstein JA, et al. Lineage structure of the human antibody repertoire in response to influenza vaccination. *Sci Transl Med*. 2013;5:171ra19.
101. Martin V, Bryan Wu YC, Kipling D, Dunn-Walters D. Ageing of the B-cell repertoire. *Philos Trans R Soc Lond B Biol Sci*. 2015;370:20140237.
102. Galson JD, Trück J, Fowler A, et al. Analysis of B cell repertoire dynamics following hepatitis B vaccination in humans, and enrichment of vaccine-specific antibody sequences. *EBioMedicine*. 2015;2:2070-2079.
103. Galson JD, Clutterbuck EA, Trück J, et al. BCR repertoire sequencing: Different patterns of B-cell activation after two Meningococcal vaccines. *Immunol Cell Biol*. 2015;93:885-895.
104. Galson JD, Trück J, Clutterbuck EA, et al. B-cell repertoire dynamics after sequential hepatitis B vaccination and evidence for cross-reactive B-cell activation. *Genome Med*. 2016;8:68.
105. Scheid JF, Mouquet H, Feldhahn N, et al. Broad diversity of neutralizing antibodies isolated from memory B cells in HIV-infected individuals. *Nature*. 2009;458:636-640.
106. Yoon H, Macke J, West AP Jr, et al. CATNAP: A tool to compile, analyze and tally neutralizing antibody panels. *Nucleic Acids Res*. 2015;43:W213-W219.
107. Asti L, Uguzzoni G, Marcatili P, Pagnani A. Maximum-entropy models of sequenced immune repertoires predict antigen-antibody affinity. *PLoS Comput Biol*. 2016;12:e1004870.
108. Galson JD, Trück J, Fowler A, et al. In-depth assessment of within-individual and inter-individual variation in the B cell receptor repertoire. *Front Immunol*. 2015;6:531.
109. Henry Dunand CJ, Wilson PC. Restricted, canonical, stereo-typed and convergent immunoglobulin responses. *Philos Trans R Soc Lond B Biol Sci*. 2015;370:20140238.
110. Trück J, Ramasamy MN, Galson JD, et al. Identification of antigen-specific B cell receptor sequences using public repertoire analysis. *J Immunol*. 2015;194:252-261.

111. Wang C, Liu Y, Cavanagh MM, et al. B-cell repertoire responses to varicella-zoster vaccination in human identical twins. *Proc Natl Acad Sci USA*. 2015;112:500-505.
112. Greiff V, Menzel U, Miho E, et al. Systems analysis reveals high genetic and antigen-driven pre-determination of antibody repertoires throughout B cell development. *Cell Rep*. 2017;19:1467-1478.
113. Collins AM, Jackson KJL. On being the right size: Antibody repertoire formation in the mouse and human. *Immunogenetics*. 2017;70:143-158.
114. DeWitt WS, Lindau P, Snyder TM, et al. A public database of memory and naive B-cell receptor sequences. *PLoS ONE*. 2016;11:e0160853.
115. DeWitt W, Lindau P, Snyder T, Vignali M, Emerson R, Robins H. Replicate immunosequencing as a robust probe of B cell repertoire diversity. *arXiv*. 2014; arXiv:1410.0350.
116. Methot SP, Di Noia JM. Chapter two-molecular mechanisms of somatic hypermutation and class switch recombination. In: Alt FW, ed. *Advances in Immunology (Vol 133)*. Cambridge, UK: Academic Press; 2017:37-87.
117. Hwang JK, Wang C, Du Z, et al. Sequence intrinsic somatic mutation mechanisms contribute to affinity maturation of VRC01-class HIV-1 broadly neutralizing antibodies. *Proc Natl Acad Sci USA*. 2017;114:8614-8619.
118. Dunn-Walters DK, Dogan A, Boursier L, MacDonald CM, Spencer J. Base-specific sequences that bias somatic hypermutation deduced by analysis of out-of-frame human IgVH genes. *J Immunol*. 1998;160:2360-2364.
119. Cowell LG, Kepler TB. The nucleotide-replacement spectrum under somatic hypermutation exhibits microsequence dependence that is strand-symmetric and distinct from that under germline mutation. *J Immunol*. 2000;164:1971-1976.
120. Pham P, Bransteitter R, Petruska J, Goodman MF. Processive AID-catalysed cytosine deamination on single-stranded DNA simulates somatic hypermutation. *Nature*. 2003;424:103-107.
121. Cui A, Di Niro R, Vander Heiden JA, et al. A model of somatic hypermutation targeting in mice based on high-throughput Ig sequencing data. *J Immunol*. 2016;197:3566-3574.
122. Cohen RM, Kleinstein SH, Louzoun Y. Somatic hypermutation targeting is influenced by location within the immunoglobulin V region. *Mol Immunol*. 2011;48:1477-1483.
123. Sheng Z, Schramm CA, Kong R, NISC Comparative Sequencing Program, Mullikin JC, Mascola JR, et al. Gene-specific substitution profiles describe the types and frequencies of amino acid changes during antibody somatic hypermutation. *Front Immunol*. 2017;8:537.
124. Kirik U, Persson H, Levander F, Greiff L, Ohlin M. Antibody heavy chain variable domains of different germline gene origins diversify through different paths. *Front Immunol*. 2017;8:1433.
125. Dhar A, Davidsen K, Matsen FA IV, Minin VN. Predicting B cell receptor substitution profiles using public repertoire data. *arXiv*. 2018;arXiv:1802.06406.
126. Dunn-Walters DK, Spencer J. Strong intrinsic biases towards mutation and conservation of bases in human IgVH genes during somatic hypermutation prevent statistical analysis of antigen selection. *Immunology*. 1998;95:339-345.
127. Yaari G, Uduman M, Kleinstein SH. Quantifying selection in high-throughput immunoglobulin sequencing data sets. *Nucleic Acids Res*. 2012;40:e134.
128. McCoy CO, Bedford T, Minin VN, Bradley P, Robins H, Matsen FA IV. Quantifying evolutionary constraints on B-cell affinity maturation. *Philos Trans R Soc Lond B Biol Sci*. 2015;370:20140244.
129. Vieira MC, Zinder D, Cobey S. Selection and neutral mutations drive pervasive mutability losses in long-lived anti-HIV B cell lineages. *Mol Biol Evol*. 2018;35:1135-1146.
130. Spencer J, Dunn M, Dunn-Walters DK. Characteristics of sequences around individual nucleotide substitutions in IgVH genes suggest different GC and AT mutators. *J Immunol*. 1999;162:6596-6601.
131. Rogozin IB, Pavlov YI, Bebenek K, Matsuda T, Kunkel TA. Somatic mutation hotspots correlate with DNA polymerase η error spectrum. *Nat Immunol*. 2001;2:530-536.
132. Wilson TM, Vaisman A, Martomo SA, et al. MSH2–MSH6 stimulates DNA polymerase η, suggesting a role for A: T mutations in antibody genes. *J Exp Med*. 2005;201:637-645.
133. Wang M, Rada C, Neuberger MS. Altering the spectrum of immunoglobulin V gene somatic hypermutation by modifying the active site of AID. *J Exp Med*. 2010;207:141-153.
134. Mak CH, Pham P, Afif SA, Goodman MF. A mathematical model for scanning and catalysis on single-stranded DNA, illustrated with activation-induced deoxycytidine deaminase. *J Biol Chem*. 2013;288:29786-29795.
135. Chahwan R, Edelmann W, Scharff MD, Roa S. AID-ing antibody diversity by error-prone mismatch repair. *Semin Immunol*. 2012;24:293-300.
136. Wilson PC, de Bouteiller O, Liu YJ, et al. Somatic hypermutation introduces insertions and deletions into immunoglobulin V genes. *J Exp Med*. 1998;187:59-70.
137. Briney BS, Willis JR, Crowe JE Jr. Location and length distribution of somatic hypermutation-associated DNA insertions and deletions reveals regions of antibody structural plasticity. *Genes Immun*. 2012;13:523-529.
138. Bowers PM, Verdino P, Wang Z, et al. Nucleotide insertions and deletions complement point mutations to massively expand the diversity created by somatic hypermutation of antibodies. *J Biol Chem*. 2014;289:33557-33567.
139. Yeap LS, Hwang JK, Du Z, et al. Sequence-intrinsic mechanisms that target AID mutational outcomes on antibody genes. *Cell*. 2015;163:1124-1137.
140. Kepler TB, Liao HX, Alam SM, et al. Immunoglobulin gene insertions and deletions in the affinity maturation of HIV-1 broadly reactive neutralizing antibodies. *Cell Host Microbe*. 2014;16:304-313.
141. Hwang DG, Green P. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proceedings of the National Academy of Sciences USA*. 2004;101:13994-14001.
142. Hobolth A. A Markov chain Monte Carlo expectation maximization algorithm for statistical analysis of DNA sequence evolution with neighbor-dependent substitution rates. *Journal of Computational and Graphical Statistics*. 2008;17:138-162.
143. Bishop MJ, Thompson EA. Maximum likelihood alignment of DNA sequences. *J Mol Biol*. 1986;190:159-165.
144. Thorne JL, Kishino H, Felsenstein J. An evolutionary model for maximum likelihood alignment of DNA sequences. *J Mol Evol*. 1991;33:114-124.
145. Thorne JL, Kishino H, Felsenstein J. Inching toward reality: An improved likelihood model of sequence evolution. *J Mol Evol*. 1992;34:3-16.
146. Bouchard-Côte′ A, Jordan MI. Evolutionary inference via the Poisson Indel Process. *Proc Natl Acad Sci USA*. 2013;110:1160-1166.
147. Zhai Y, Alexandre BC. A Poissonian model of indel rate variation for phylogenetic tree inference. *Syst Biol*. 2017;66:698-714.
148. Mayer CT, Gazumyan A, Kara EE, et al. The microanatomic segregation of selection by apoptosis in the germinal center. *Science*. 2017;358:eaao2602.
149. Tas JMJ, Mesin L, Pasqual G, et al. Visualizing antibody affinity maturation in germinal centers. *Science*. 2016;351:1048-1054.
150. Kuraoka M, Schmidt AG, Nojima T, et al. Complex antigens drive permissive clonal selection in germinal centers. *Immunity*. 2016;44:542-552.

151. Zhang Y, Meyer-Hermann M, George LA, et al. Germinal center B cells govern their own fate via antibody feedback. *J Exp Med*. 2013;210:457-464.
152. de Bourcy CFA, Angel CJL, Vollmers C, Dekker CL, Davis MM, Quake SR. Phylogenetic analysis of the human antibody repertoire reveals quantitative signatures of immune senescence and aging. *Proc Natl Acad Sci USA*. 2017;114:1105-1110.
153. Or-Guil M, Wittenbrink N, Weiser AA, Schuchhardt J. Recirculation of germinal center B cells: A multilevel selection strategy for antibody maturation. *Immunol Rev*. 2007;216:130-141.
154. McHeyzer-Williams LJ, Milpied PJ, Okitsu SL, McHeyzer-Williams MG. Class-switched memory B cells remodel BCRs within secondary germinal centers. *Nat Immunol*. 2015;16:296-305.
155. Gupta NT, Adams KD, Briggs AW, Timberlake SC, Vigneault F, Kleinstein SH. Hierarchical clustering can identify B cell clones with high confidence in Ig repertoire sequencing data. *J Immunol*. 2017;198:2489-2499.
156. Wu X, Zhou T, Zhu J, et al. Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing. *Science*. 2011;333:1593-1602.
157. Saitou N, Nei M. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol Biol Evol*. 1987;4:406-425.
158. Gascuel O, Steel M. Neighbor-joining revealed. *Mol Biol Evol*. 2006;23:1997-2000.
159. Doucet A, de Freitas N, Gordon N. An introduction to sequential Monte Carlo methods. In: Doucet A, de Freitas N, Gordon N, eds. *Sequential Monte Carlo Methods in Practice. Statistics for Engineering and Information Science*. New York, NY: Springer; 2001:3-14.
160. Laserson J. *Bayesian assembly of reads from high throughput sequencing*. Doctoral dissertation, Stanford University; 2012.
161. Barak M, Zuckerman N, Edelman H, Unger R, Mehr R. IgTree (c): Creating immunoglobulin variable region gene lineage trees. *J Immunol Methods*. 2008;338:67-74.
162. Stern JNH, Yaari G, Vander Heiden JA, et al. B cells populating the multiple sclerosis brain mature in the draining cervical lymph nodes. *Sci Transl Med*. 2014;6:248ra107.
163. Felsenstein J. Cases in which parsimony or compatibility methods will be positively misleading. *Syst Zool*. 1978;27:401-410.
164. DeWitt WS, Mesin L, Victora GD, Minin VN, Matsen FA IV. Using genotype abundance to improve phylogenetic inference. *Mol Biol Evol*. 2018;35:1253-1265.
165. Kepler TB. Reconstructing a B-cell clonal lineage. I. Statistical inference of unobserved ancestors. *F1000Res*. 2013;2:103.
166. Hoehn KB, Lunter G, Pybus OG. A phylogenetic codon substitution model for antibody lineages. *Genetics*. 2017;206:417-427.
167. Zhu J, Wu X, Zhang B, et al. De novo identification of VRC01 class HIV-1–neutralizing antibodies by next-generation sequencing of B-cell transcripts. *Proceedings of the National Academy of Sciences of the USA*. 2013;110:E4088-E4097.
168. Doria-Rose NA, Schramm CA, Gorman J, et al. Developmental pathway for potent V1V2-directed HIV-neutralizing antibodies. *Nature*. 2014;509:55-62.
169. Felsenstein J. PHYLIP-phylogeny inference package (version 3.2). *Cladistics*. 1989;5:164-166.
170. Ashkenazy H, Penn O, Doron-Faigenboim A, et al. FastML: A web server for probabilistic reconstruction of ancestral sequences. *Nucleic Acids Res*. 2012;40:W580-W584.
171. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*. 2015;32:268-274.
172. Hershberg U, Uduman M, Shlomchik MJ, Kleinstein SH. Improved methods for detecting selection by mutation analysis of Ig V region sequences. *Int Immunol*. 2008;20:683-694.
173. Uduman M, Yaari G, Hershberg U, Stern JA, Shlomchik MJ, Kleinstein SH. Detecting selection in immunoglobulin sequences. *Nucleic Acids Res*. 2011;39:W499-W504.
174. Yaari G, Benichou JIC, Vander Heiden JA, Kleinstein SH, Louzoun Y. The mutation patterns in B-cell immunoglobulin receptors reflect the influence of selection acting at multiple time-scales. *Philos Trans R Soc Lond B Biol Sci*. 2015;370:20140242.
175. Dunn-Walters DK, Belelovsky A, Edelman H, Banerjee M, Mehr R. The dynamics of germinal centre selection as measured by graph-theoretical analysis of mutational lineage trees. *Dev Immunol*. 2002;9:233-243.
176. Uduman M, Shlomchik MJ, Vigneault F, Church GM, Kleinstein SH. Integrating B cell lineage information into statistical tests for detecting selection in Ig sequences. *J Immunol*. 2014;192:867-874.
177. Liberman G, Benichou JIC, Maman Y, Glanville J, Alter I, Louzoun Y. Estimate of within population incremental selection through branch imbalance in lineage trees. *Nucleic Acids Res*. 2016;44:e46.
178. Horns F, Vollmers C, Dekker CL, Quake SR. Signatures of selection in the human antibody repertoire: Selective sweeps, competing subclones, and neutral drift. *BioRxiv*. 2017. https://doi.org/10.1101/145052.
179. Neher RA, Russell CA, Shraiman BI. Predicting evolution from the shape of genealogical trees. *Elife*. 2014;3:e03568.
180. Fay JC, Wu CI. Hitchhiking under positive Darwinian selection. *Genetics*. 2000;155:1405-1413.
181. Shlomchik MJ, Watts P, Weigert MG, Litwin S. Clone: A Monte-Carlo computer simulation of B cell clonal expansion, somatic mutation, and antigen-driven selection. In: Kelsoe G, Flajnik PMF, eds. *Somatic Diversification of Immune Responses. Current Topics in Microbiology and Immunology*. Berlin, Heidelberg: Springer; 1998:173-197.
182. Kim PS, Levy D, Lee PP. Modeling and simulation of the immune system as a self-regulating network. In: Johnson ML, Brand L, eds. *Methods in Enzymology*. Vol. 467. Cambridge, UK: Academic Press; 2009:79-109.
183. Kleinstein SH, Singh JP. Toward quantitative simulation of germinal center dynamics: Biological and modeling insights from experimental validation. *J Theor Biol*. 2001;211:253-275.
184. Mehr R. Asynchronous differentiation models explain bone marrow labeling kinetics and predict reflux between the pre-and immature B cell pools. *Int Immunol*. 2003;15:301-312.
185. Shahaf G, Allman D, Cancro MP, Mehr R. Screening of alternative models for transitional B cell maturation. *Int Immunol*. 2004;16:1081-1090.
186. Shahaf G, Johnson K, Mehr R. B cell development in aging mice: Lessons from mathematical modeling. *Int Immunol*. 2006;18:31-39.
187. Shahaf G, Cancro MP, Mehr R. Kinetic modeling reveals a common death niche for newly formed and mature B cells. *PLoS ONE*. 2010;5:e9497.
188. Childs LM, Baskerville EB, Cobey S. Trade-offs in antibody repertoires to complex antigens. *Philos Trans R Soc Lond B Biol Sci*. 2015;370:20140245.
189. Wang S, Mata-Fink J, Kriegsman B, et al. Manipulating the selection forces during affinity maturation to generate cross-reactive HIV antibodies. *Cell*. 2015;160:785-797.
190. Wang S. Optimal sequential immunization can focus antibody responses against diversity loss and distraction. *PLoS Comput Biol*. 2017;13:e1005336.
191. Amitai A, Mesin L, Victora G, Kardar M, Chakraborty A. A population dynamics model for clonal diversity in a germinal center. *Front Microbiol*. 2017;8:1693.
192. Liao HX, Lynch R, Zhou T, et al. Coevolution of a broadly neutralizing HIV-1 antibody and founder virus. *Nature*. 2013;496:469-476.

193. Luo S, Perelson AS. The challenges of modelling antibody repertoire dynamics in HIV infection. *Philos Trans R Soc Lond B Biol Sci*. 2015;370:20140247.
194. Hershberg U, Luning Prak ET. The analysis of clonal expansions in normal and autoimmune B cell repertoires. *Philos Trans R Soc Lond B Biol Sci*. 2015;370:20140239.
195. Sok D, Laserson U, Laserson J, et al. The effects of somatic hypermutation on neutralization and binding in the PGT121 family of broadly neutralizing HIV antibodies. *PLoS Pathog*. 2013;9:e1003754.
196. de Koning APJ, Gu W, Pollock DD. Rapid likelihood analysis on large phylogenies using partial sampling of substitution histories. *Mol Biol Evol*. 2010;27:249-265.
197. Chib S. Marginal likelihood from the gibbs output. *J Am Stat Assoc*. 1995;90:1313-1321.
198. Drummond AJ, Rambaut A, Shapiro B, Pybus OG. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol*. 2005;22:1185-1192.
199. Perelson AS, Oster GF. Theoretical studies of clonal selection: Minimal antibody repertoire size and reliability of self-non-self discrimination. *J Theor Biol*. 1979;81:645-670.
200. Hwang JK, Alt FW, Yeap LS. Related mechanisms of antibody somatic hypermutation and class switch recombination. *Microbiol Spectr*. 2015;3:MDNA3–0037–2014.
201. Jackson KJL, Wang Y, Collins AM. Human immunoglobulin classes and subclasses show variability in VDJ gene mutation levels. *Immunol Cell Biol*. 2014;92:729-733.
202. Horns F, Vollmers C, Croote D, et al. Lineage tracing of human B cells reveals the in vivo landscape of human antibody class switching. *Elife*. 2016;5:e16578.
203. Looney TJ, Lee JY, Roskin KM, et al. Human B-cell isotype switching origins of IgE. *J Allergy Clin Immunol*. 2016;137:579-586 e7.
204. Pagel M. Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. *Proc R Soc Lond B Biol Sci*. 1994;255:37-45.
205. Pagel M, Meade A, Barker D. Bayesian estimation of ancestral character states on phylogenies. *Syst Biol*. 2004;53:673-684.
206. Kaplinsky J, Arnaout R. Robust estimates of overall immunerepertoire diversity from high-throughput measurements on samples. *Nat Commun*. 2016;7:11881.
207. Meng W, Zhang B, Schwartz GW, et al. An atlas of B-cell clonal distribution in the human body. *Nat Biotechnol*. 2017;35:879-884.
208. Havenar-Daughton C, Carnathan DG, Torrents de la Peña A, et al. Direct probing of germinal center responses reveals immunological features and bottlenecks for neutralizing antibody responses to HIV Env trimer. *Cell Rep*. 2016;17:2195-2209.
209. Wine Y, Boutz DR, Lavinder JJ, et al. Molecular deconvolution of the monoclonal antibodies that comprise the polyclonal serum response. *Proc Natl Acad Sci USA*. 2013;110:2993-2998.
210. Lavinder JJ, Wine Y, Giesecke C, et al. Identification and characterization of the constituent human serum antibodies elicited by vaccination. *Proc Natl Acad Sci USA*. 2014;111:2259-2264.
211. Nivala J, Marks DB, Akeson M. Unfoldase-mediated protein translocation through an α-hemolysin nanopore. *Nat Biotechnol*. 2013;31:247–250.
212. Doolan DL, Mu Y, Unal B, et al. Profiling humoral immune responses to P. falciparum infection with protein microarrays. *Proteomics*. 2008;8:4680-4694.
213. Hertz T, Beatty PR, MacMillen Z, Killingbeck SS, Wang C, Harris E. Antibody epitopes identified in critical regions of dengue virus non-structural 1 protein in mouse vaccination and natural human infections. *J Immunol*. 2017;198:4025-4035.
214. Gee MH, Han A, Lofgren SM, et al. Antigen identification for orphan T cell receptors expressed on tumor-infiltrating lymphocytes. *Cell*. 2017;172:549-563 e16.
215. Smith DJ, Lapedes AS, de Jong JC, et al. Mapping the antigenic and genetic evolution of influenza virus. *Science*. 2004;305:371-376.
216. Katzelnick LC, Fonville JM, Gromowski GD, et al. Dengue viruses cluster antigenically but not as discrete serotypes. *Science*. 2015;349:1338-1343.
217. Francis T Jr. On the doctrine of original antigenic sin. *Proc Am Philos Soc*. 1960;104:572-578.
218. Andrews SF, Huang Y, Kaur K, et al. Immune history profoundly affects broadly protective B cell responses to influenza. *Sci Transl Med*. 2015;7:316ra192.
219. Horwitz JA, Bar-On Y, Lu CL, et al. Non-neutralizing antibodies alter the course of HIV-1 infection in vivo. *Cell*. 2017;170:637-648 e10.
220. Łuksza M, Lässig M. A predictive fitness model for influenza. *Nature*. 2014;507:57-61.
221. Neher RA, Bedford T, Daniels RS, Russell CA, Shraiman BI. Prediction, dynamics, and visualization of antigenic phenotypes of seasonal influenza viruses. *Proc Natl Acad Sci USA*. 2016;113:E1701-E1709.
222. Zarnitsyna VI, Ellebedy AH, Davis C, Jacob J, Ahmed R, Antia R. Masking of antigenic epitopes by antibodies shapes the humoral immune response to influenza. *Philos Trans R Soc Lond B Biol Sci*. 2015;370:20140248.
223. Zarnitsyna VI, Lavine J, Ellebedy A, Ahmed R, Antia R. Multi-epitope models explain how pre-existing antibodies affect the generation of broadly protective responses to influenza. *PLoS Pathog*. 2016;12:e1005692.
224. Doud MB, Hensley SE, Bloom JD. Complete mapping of viral escape from neutralizing antibodies. *PLoS Pathog*. 2017;13:e1006271.
225. Dingens AS, Haddox HK, Overbaugh J, Bloom JD. Comprehensive mapping of HIV-1 escape from a broadly neutralizing antibody. *Cell Host Microbe*. 2017;21:777-787 e4.
226. Doud MB, Lee JM, Bloom JD. How single mutations affect viral escape from broad and narrow antibodies to H1 influenza hemagglutinin. *Natu Commun*. 2018;9:1386.
227. Forsyth CM, Juan V, Akamatsu Y, et al. Deep mutational scanning of an antibody against epidermal growth factor receptor using mammalian cell display and massively parallel pyrosequencing. *MAbs*. 2013;5:523-532.
228. Adams RM, Mora T, Walczak AM, Kinney JB. Measuring the sequence-affinity landscape of antibodies with massively parallel titration curves. *Elife*. 2016;5:e23156.
229. Fonville JM, Wilks SH, James SL, et al. Antibody landscapes after influenza virus infection or vaccination. *Science*. 2014;346:996-1000.
230. Brodin P, Davis MM. Human immune system variation. *Nat Rev Immunol*. 2017;17:21-29.

How to cite this article: Olson BJ, Matsen FA IV. The Bayesian optimist's guide to adaptive immune receptor repertoire analysis. *Immunol Rev*. 2018;284:148–166.
https://doi.org/10.1111/imr.12664