Per-sample immunoglobulin germline inference from B cell receptor deep sequencing data

Duncan K. Ralph[1,*], Frederick A. Matsen IV[1]

1 Fred Hutch, Seattle, Washington, USA

[*] **dkralph@gmail.com**

Abstract

The collection of immunoglobulin genes in an individual's germline, which gives rise to B cell receptors via recombination, is known to vary significantly across individuals. In humans, for example, each individual has only a fraction of the several hundred known V alleles. Furthermore, this set of known V alleles is both incomplete (particularly for non-European samples), and contains a significant number of spurious alleles. The resulting uncertainty as to which immunoglobulin alleles are present in any given sample results in inaccurate B cell receptor sequence annotations, and in particular inaccurate inferred naive ancestors. In this paper we first show that the currently widespread practice of aligning each sequence to its closest match in the full set of IMGT alleles results in a very large number of spurious alleles that are not in the sample's true set of germline V alleles. We then describe a new method for inferring each individual's germline gene set from deep sequencing data, and show that it improves upon existing methods by making a detailed comparison on a variety of simulated and real data samples. This new method has been integrated into the partis annotation and clonal family inference package, available at `https://github.com/psathyrella/partis`, and is run by default without affecting overall run time.

Author Summary

Antibodies are an important component of the adaptive immune system, which itself determines our response to both pathogens and vaccines. They are produced by B cells through somatic recombination of germline DNA, which results in a vast diversity of antigen binding affinities across the B cell repertoire. We typically learn about the development of this repertoire, and its history of interaction with antigens, by sequencing large numbers of the DNA sequences from which antibodies are derived. In order to understand such data, it is necessary to determine the combination of germline V, D, and J genes that was rearranged to form each such B cell receptor sequence. This is difficult, however, because the immunoglobulin locus exhibits an extraordinary level of diversity across individuals – encompassing both allelic variation and gene duplication, deletion, and conversion – and because the locus's large size and repetitive structure make germline sequencing very difficult. In this paper we describe a new computational method that avoids this difficulty by inferring each individual's set of immunoglobulin germline genes directly from expressed B cell receptor sequence data.

Introduction

The heavy and light chain B cell receptor (BCR) loci arise from a random recombination of germline V, D, and J genes. Repeated across many B cells, this generates the vast diversity of naive BCRs that is integral to the adaptive immune system. As an additional source of population-wide variation, there is significant variation of germline genes between individuals. Databases such as IMGT [1] aim to collect and organize this ensemble of germline genes.

The analysis of BCR sequence data begins with the alignment of each sequence against a set of germline V, D, and J genes. A variety of methods (e.g. [2–5]) have been developed to accomplish the basic task of deciding which V, D, and J genes gave rise to each observed sequence. There has been less work, however, toward measuring the extent to which the set of germline genes used for this analysis resembles the germline gene set actually present in the individual from which the sequence data was derived. Most methods simply use the full set of germline genes from a database such as IMGT [1] for all samples.

One problem with this approach is that the IMGT set includes genes from all individuals of a species, while any single individual's germline contains only a fraction of these (roughly 50 out of 250 V genes, 25 of 35 D, and 6 of 12 J). This is problematic for sequencing studies that use antigen-experienced B cells that have been through several rounds of somatic hypermutation (SHM), which obscures the identity of the original germline gene. As we show below, this leads to large numbers of spurious gene assignments, and an inferred germline gene set with many more alleles than are in the individual's true set.

Another problem with this approach is that no database contains a perfect catalog of the complete immunoglobulin germline diversity of each species. Sequencing continues to uncover novel human V genes that are not in any previous database [6–13]. Additionally, a significant fraction of the sequences in existing databases are likely the result of sequencing error rather than real biological variation [14–16]. Our knowledge of the immunoglobulin locus is even less complete for other species [12, 17].

Improving our understanding of the immunoglobulin locus, however, is not simply a matter of applying standard genome sequencing protocols more broadly. Most genome sequencing is performed on lymphoblastoid cell lines [18–20], whose prior rearrangement has destroyed much of the information about the original immunoglobulin locus. The obvious solution would be to sequence other cell types; however assembly challenges due to the complexity and repetitiveness of the locus [21] mean that even sequencing an intact immunoglobulin locus is not straightforward. The IGHV locus, for instance, consists of about 120 V genes, roughly two-thirds of which are non-functional pseudogenes, spread over a megabase of chromosome 14 [9]. The immunoglobulin locus is also subject to widespread gene duplication, deletion, and conversion [7, 8, 22, 23]. Thus although databases such as the 1000 Genomes project and the Simons Genome Diversity Project can be used to investigate immunoglobulin diversity [23, 24], this approach is not without pitfalls [25].

Discrepancies between a BCR-sequenced individual's true set of germline genes and the set used to analyze their BCR sequences cause a number of practical problems. First, finding associations between particular germline genes and an immunological response is difficult if the gene assignment itself is suspect. This would impact, for example, recent work on the effects of the presence or absence of individual alleles on broadly neutralizing anti-influenza antibody development [26]. Second, such misassignment leads to inaccurate inferred naive ancestor sequences. Efforts to synthesize these inaccurate ancestral sequences in the lab and study their binding properties may then result in erroneous conclusions, since even single amino acid changes can have large effects on affinity [27]. And finally, studies of mutation [28, 29] and selection [30, 31] during affinity maturation depend upon accurate inferred naive sequences in order to correctly identify somatic mutations.

Our current understanding of the immunoglobulin locus comes largely from a small number of low-throughput genome and BAC library sequencing studies. The first complete sequence of the locus [32], which has been included in the first few drafts of the human genome, was assembled from several different cell lines and is therefore not a haplotype. More recently, a single complete haplotype of the heavy [9] and light [10] chain loci has been published. In addition to these larger efforts, many less-comprehensive studies of the locus have been cataloged at `www.imgt.org`.

Advances in sequencing technology, however, have allowed progress to come also from inference on expressed BCR repertoires. Several initial studies inferred germline sets by combining computational analysis with expert scrutiny, with one paper reporting a high level of diversity with many novel (non-IMGT) alleles across 12 individuals [7], and a second extending those results to 18 complete haplotypes [8]. Similar work by a different team used naive sequences to infer germline sets and haplotype linkage information for two individuals [33]. None of these studies, however, resulted in a generally-applicable software package or included a broad-scale validation of their methods.

More recently, software packages have been developed that enable fully-automated germline inference including novel allele discovery. TIgGER [11] uses a detailed per-position fitting procedure to find new alleles separated by a small number of point mutations from genes in a known database, and a heuristic prevalence threshold-based procedure to infer germline sets. The IgDiscover package [12] infers germline sets using Levenshtein distance-based hierarchical UPGMA clustering on low-SHM IgM samples. This approach allows IgDiscover to find new alleles separated by an arbitrary number of point mutations and insertion/deletion events, and frees it from the need for an initial species-specific starting database.

In this paper we present a new method for automated inference of per-sample germline V gene sets from expressed BCR sequence data. We first compare our method's accuracy on a variety of simulated samples both to the common practice of aligning against the full IMGT set, and to the two existing germline inference methods, TIgGER and IgDiscover. We find that use of the full IMGT set results in a very large number of spuriously-inferred alleles on typical samples, as well as inaccurately inferred naive sequences. We further find that while our method infers a similar fraction of correct and incorrect genes as TIgGER and IgDiscover, its inferred genes are more similar to the true genes, and thus our method's inferred naive sequences are significantly more accurate. We then use a variety of real data samples from the literature to compare the germline gene sets inferred by our method to those from TIgGER and IgDiscover, and find a similar level of concordance as in simulation. Because our method performs well on samples with elevated levels of SHM, it is more generally applicable than IgDiscover, which is restricted to low-SHM IgM samples. In addition, while TIgGER and IgDiscover are essentially standalone germline set inference packages, our method is integrated with and run by default in the general-purpose partis package, which also provides simulation, annotation, and clonal family inference. Because D inference would be very challenging, and because the J locus varies much less between individuals than either V or D, in this paper we follow these other software packages in limiting ourselves to studies of V diversity.

Because of the high prevalence of both single nucleotide polymorphisms (SNPs) and structural variants in the immunoglobulin locus, there is no single reference genome to which all variants can be mapped, and thus standard SNP nomenclature appears insufficient. In this paper the usage of "gene" and "allele" is thus largely interchangeable. In addition, we define the "germline haplotype" as the set of germline genes on a single chromosome, while "germline gene set" refers to the full set on both the maternal and paternal chromosomes. In cases where confusion is unlikely, the latter will be shortened to "germline set".

Results

Simulation methods summary. In order to establish an expectation for how germline inference methods will perform on real data, we first investigate performance on a number of simulation samples. BCR repertoires differ significantly in many different variables such as SHM levels, germline set complexity, and clonal family structure. Although we would in principle like to explore germline inference accuracy by varying all of these variables simultaneously, this is combinatorially infeasible, and we thus adopt a two-stage approach to validation. We first vary one variable at a time, while holding all others constant, using simplified "sparse" repertoires consisting of sequences stemming from only a few genes. We then choose several representative values for each variable, and simulate full, realistic repertoires at these values. Geometrically, this can be imagined as investigating performance first along many slices through the parameter space, and then at several fixed points. This approach is motivated by the fact that, in sequence-similarity space, realistic repertoires are composed of widely-spaced groups of genes, where each group consists of a few genes that are much closer to each other than the typical between-group spacing. The genes within each group are thus easily confused with each other due to SHM, but not with genes in other groups. The sparse repertoires effectively recreate the dynamics within such a group, while allowing exploration of a much larger portion of parameter space than if we were to use full repertoires for all simulations.

In this paper, the germline set for each sparse repertoire consists of one known germline gene, and either one or two novel alleles. Each full-repertoire sample, meanwhile, is generated by choosing a number of V, D, and J genes, and some number of alleles for each of these genes, based on results from germline sequencing studies (see Methods), which results in roughly 55 V, 25 D, and 6 J alleles per sample.

Validation results.

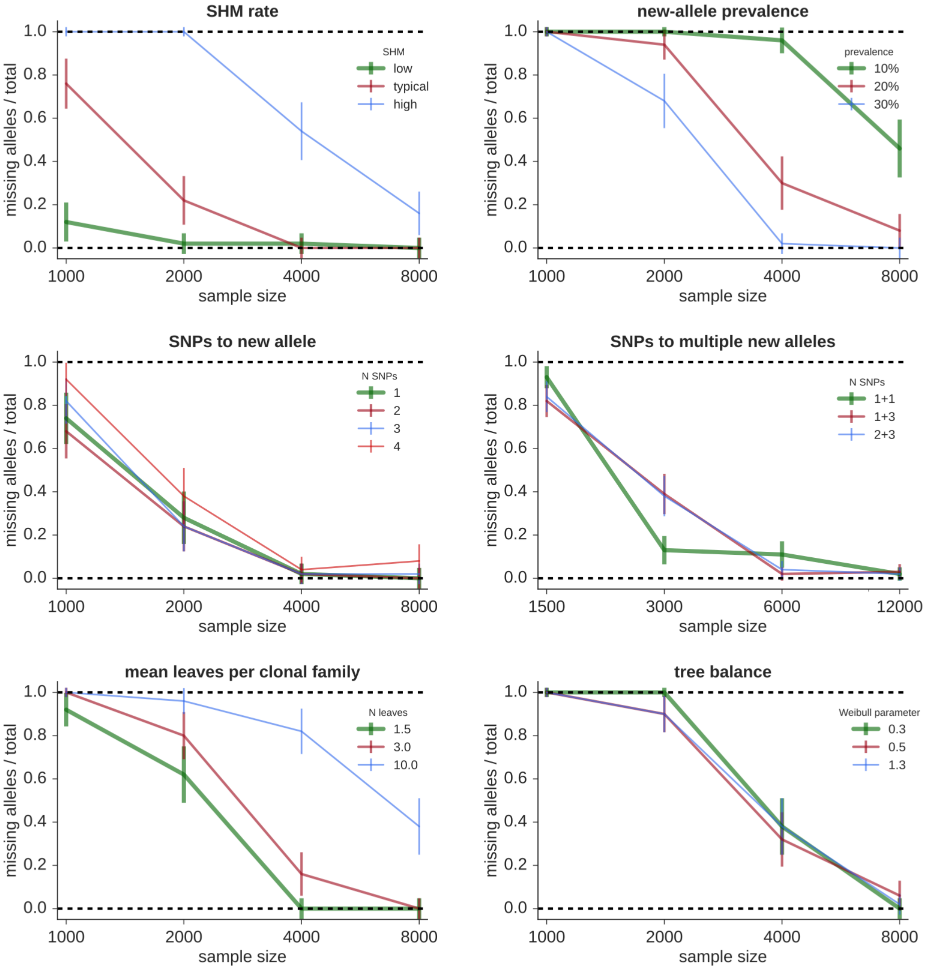

FIGURE 1. **Fraction of true alleles missing** (not inferred) by partis on simplified "sparse" repertoires for a variety of variables as a function of the number of sequences in the sample. **Top left:** SHM levels (the SHM distributions corresponding to "low", "typical", and "high" are shown in Fig S1). **Top right:** new-allele prevalence (as a fraction of the existing allele's prevalence). **Middle left:** number of SNPs (N_{snp}) separating new and existing alleles. **Middle right:** N_{snp} with multiple new alleles, where, e.g. "1 + 3" indicates two new alleles, separated by 1 and 3 SNPs from the same existing allele. **Bottom left:** mean number of leaves per clonal family. **Bottom right:** tree balance. Each point represents the mean performance (± standard error) on 50 independent simulation samples of the indicated sample size.

Variation of individual variables on sparse repertoires. Using partis's germline set inference algorithm, we quantified the impact of six repertoire characteristics on sensitivity and specificity. We did so by plotting the fraction of alleles in the true

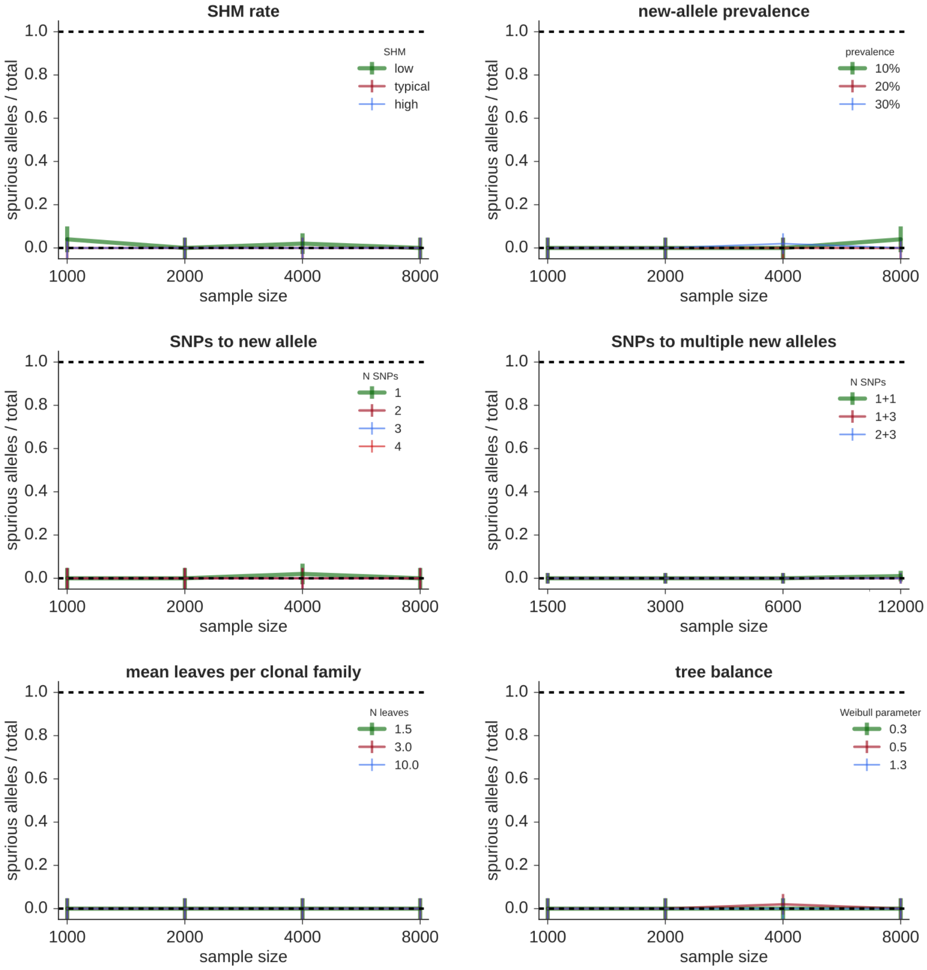

FIGURE 2. **Fraction of partis-inferred alleles not in the true germline set** on simplified "sparse" repertoires for a variety of variables as a function of the number of sequences in the sample. For explanation see Fig 1.

repertoire that are missing from the inferred repertoire (Fig 1), and the fraction of spuriously-inferred alleles (that are not in the true repertoire, Fig 2) as a function of sample size for each variable.

Increasing the rate of SHM makes inference more challenging (Figs 1 and 2, top left, with the corresponding SHM distributions in Fig S1). Because allele inference sensitivity is determined mainly by sequences with a small number of SHMs (specifically, a number comparable to the number of SNPs separating the new and existing alleles), raising SHM rates effectively reduces sample size.

Alleles that occur at low prevalence are more difficult to infer: as the fraction of sequences stemming from the new allele decreases, so does sensitivity (Figs 1 and 2, top right).

		V naive inaccuracy	# missing	# spurious	# correct
low SHM	full IMGT	0.22 ± 0.02	5.0 ± 0.0	53.0 ± 2.4	50.3 ± 0.7
	IgDiscover	0.36 ± 0.08	3.7 ± 0.2	3.4 ± 0.3	51.6 ± 0.6
	TIgGER	0.42 ± 0.07	2.7 ± 0.3	0.4 ± 0.2	52.6 ± 0.6
	partis	0.08 ± 0.02	2.4 ± 0.4	1.3 ± 0.3	52.9 ± 0.9
high SHM	full IMGT	0.31 ± 0.02	5.0 ± 0.0	80.3 ± 3.5	51.5 ± 1.0
	TIgGER	1.78 ± 0.36	7.8 ± 0.4	0.0 ± 0.0	48.7 ± 0.9
	partis	0.27 ± 0.03	9.3 ± 0.7	2.9 ± 0.3	47.2 ± 0.7

TABLE 1. **Summary full-repertoire simulation performance for the three germline inference methods plus "full IMGT" annotation**. Results are the mean (± standard error) of ten independent 50,000-sequence samples for both low-SHM (top) and high-SHM (bottom). **V naive inaccuracy** is the mean Hamming distance between true and inferred V region naive sequences (excluding the three most 3′ bases). We also show the mean number of true alleles missing from the inferred germline set (**# missing**), the number inferred that are not in the true germline set (**# spurious**), and the number in common between the inferred and true germline sets (**# correct**). We show IgDiscover only for the low-SHM samples, since it is designed only for IgM.

The number of SNPs (N_{snp}) separating a new allele from its most similar known counterpart also affects the details of germline inference. We show performance for different N_{snp} for both a single new allele (Figs 1 and 2, middle left) and for several combinations of multiple new alleles (Figs 1 and 2, middle right). Sensitivity is independent of N_{snp} for smaller N_{snp} (three or less), and then decreases slightly with increasing N_{snp}. The presence of multiple new alleles, on the other hand, does not appreciably affect sensitivity as long as their SNPs do not occur at the same positions. Because the occurrence of multiple new alleles with the same SNP positions is rare in real data, we do not show results for this case. In many cases it is in fact possible to disentangle such alleles, but this depends on the details of each new allele's prevalence and N_{snp}.

The shared mutations within a clonal family complicate allele inference because independent mutations are required for accurate fitting. We find that increasing clonality effectively decreases sample size (Figs 1 and 2, bottom left), rather than introducing the spurious alleles that would result from fitting with non-independent mutations. This indicates that our method of selecting a small number of sequences to represent each clonal family (see Methods) provides a sufficiently accurate method of choosing sequences with independent mutations.

We find that variations in phylogenetic tree shape do not greatly affect our method (Figs 1 and 2, bottom right). We change tree shape by using the TreeSimGM package [38] to vary the shape parameter of a Weibull distribution controlling an age-dependent speciation process.

These single-variable results show that our method's sensitivity is high enough to give useful results with the sample sizes and SHM rates characteristic of typical full-repertoire samples, and that it models repertoire details well enough that spurious alleles are rare. Note that TIgGER and IgDiscover are not shown on

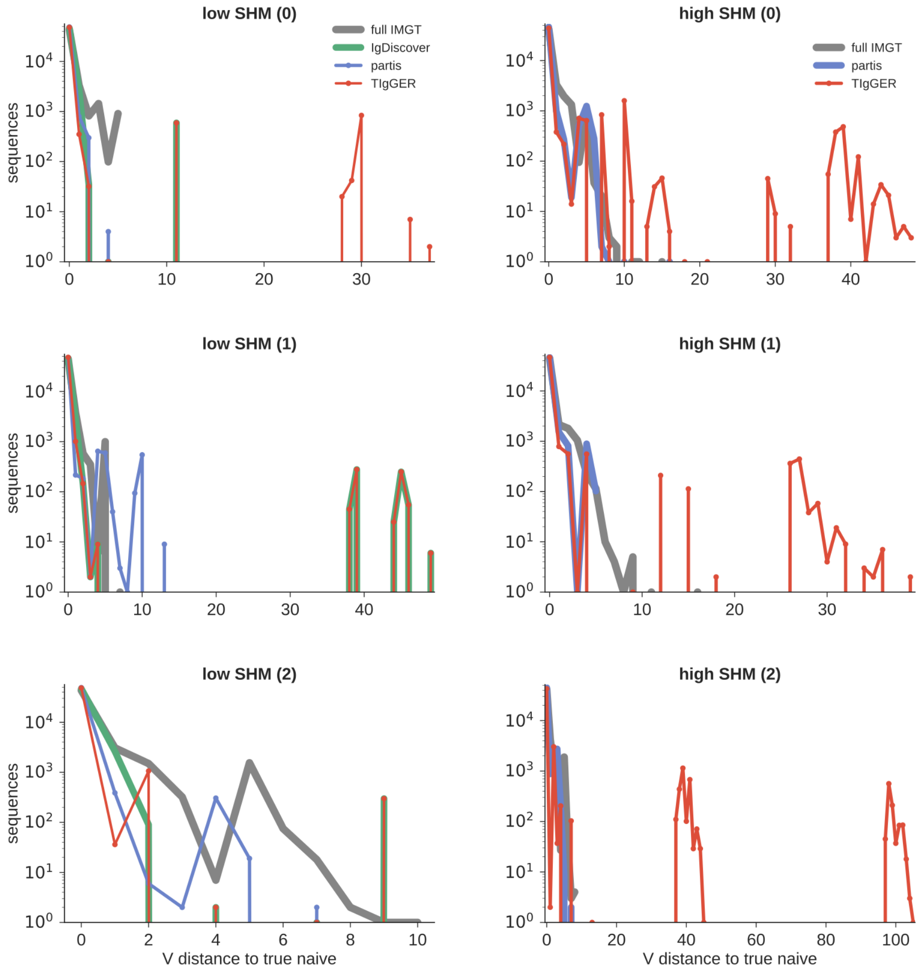

FIGURE 3. **Full-repertoire V naive accuracy** (Hamming distance between true and inferred V naive sequences) for the three germline inference methods, plus annotation with the full IMGT set. Each point represents the number of sequences (y) with a given error (Hamming distance, x). Shown on the first three replicates (0-2) of both the low-SHM (left), and high-SHM (right) full-repertoire simulation samples (see text).

these sparse samples because both methods use hard-coded assumptions tailored to typical full repertoires that cause crashes on these sparse repertoires.

Full-repertoire samples. In the second validation stage, we show performance on a smaller number of large, realistic repertoires using partis (v0.9.0), TIgGER (v0.2.10), IgDiscover (v0.6.0), and annotation with the full IMGT set. All software was run with default parameters.

We split these full-repertoire samples among two difficulty levels: ten samples with more-uniform allele prevalence and low SHM, and ten samples with less-uniform allele prevalence and higher, typical SHM (details in Methods). Results with IgDiscover are shown only for the low-SHM samples, since IgDiscover is designed to work only on low-SHM IgM-specific data.

We measure the influence of germline set accuracy on practical results in two ways: in terms of the actual genes and alleles inferred, and in terms of the resulting annotation accuracy. The former is more relevant to germline databases and studies of gene association, while the latter is of more concern when inferring and studying the function of ancestral sequences.

We find that the practice of aligning against the full IMGT set results in a very large number of spurious gene inferences, even on low-SHM samples (Table 1 and Fig 4). The three explicit germline inference methods, while all giving much smaller numbers of spurious genes, harbor significant differences. The partis-inferred missing and spurious alleles are found on relatively short branches compared to those of the other programs (Figs 5, 6, 7). This results in partis's significantly more accurate V naive inference (Table 1). By considering the distribution of Hamming distances between true and inferred naive V sequences (Fig 3), we see that the relative inaccuracy of TIgGER and IgDiscover is driven by rare sequences that are assigned to genes that are very dissimilar to their true gene. We also note that TIgGER shows reduced sensitivity at typical SHM rates (Fig 6 right), compared to low SHM rates (Fig 6 left), in fact failing to infer any of the novel (non-IMGT) alleles at typical SHM rates.

These simulation samples, together with true and inferred germline sets, are available at `https://zenodo.org/record/1037464#.WfISc3BrwUE`.

Results on real data. In order to evaluate performance on real data, it would be natural to deep sequence individuals for whom we also have accurate results from germline sequencing. Unfortunately, as described above, the difficulty of germline sequencing means that such samples are not readily available. We instead use two types of comparison that, while not definitive, provide some insight.

We first compare results from the different inference methods when run on the same sample, and find agreement on 70-90% of the total genes (Figs 8, 9, 10, S2, S3). While this is reassuring, some caution is advised, as the methods are far from uncorrelated (see Supplement). IgDiscover is shown only for IgM samples (Fig 10, and non-IgM samples with very low SHM rates (Figs 8, S2, S3). Also of note is the large cluster of closely-related novel alleles inferred only by TIgGER in the IgM data from subject lp23810 [36] (Figs 10, 13).

We next compare the results of each inference method on several different samples from the same individual. We find a similar overall level of agreement both when comparing samples from different time points (Figs 11, 12), and of different isotypes (Fig 13). These comparisons give some idea of each method's uncertainty because, while the physical germline genes are in each case identical, the SHM rates, gene expression levels, and clonal family structure vary significantly with both time and isotype.

We also use the partis-inferred germline sets to make an estimate of the number of genes that are expressed at levels too low for us to detect. Previous work has reported a range of values for the total number of functional V genes per individual. One study [9] reported 43 full-length functional V genes (plus 1 truncated) for a single haplotype, while another [39] reported a range of 38-46 per haplotype. In order to convert these per-haplotype totals to per-diplotype totals, we calculate the mean fraction of alleles shared between the inferred germline sets from two unrelated individuals. For the sequencing data in this paper, this mean overlap is 67% (range 50-85%). This suggests that to go from per-haplotype to per-diplotype totals we multiply by $1 + (1 - 0.67)$, which yields per-diplotype estimates of 57 for [9] and 51-61 for [39]. These values, both for total genes and for the fraction of genes shared between unrelated parental haplotypes, roughly agree with two other studies that found 35-46 per haplotype and 39-55 per diplotype [8], and 45-60 per diplotype with a mean alleles per gene of 1.2 [7]. The mean total number of partis-inferred V genes observed in individuals in this paper, meanwhile, is 47 (range 38-62). This suggests that the sample sizes, clonal family structures, mutation rates, and expression levels, together with our method's sensitivity, result in a failure to detect about 0 to 10 genes per individual. We have not accounted for spuriously-inferred alleles in this calculation because our validation results suggest that when partis does infer spurious alleles, each simply replaces a very similar true allele, and thus does not have an appreciable net effect.

The fasta files for each inferred germline set are available at `https://zenodo.org/record/1037464#.WfISc3BrwUE`. We have made the command-line script used to make phylogenetic comparison plots available for general application at `https://git.io/vFo2B`.

Discussion

We have developed a practical new tool for inferring per-sample immunoglobulin germline gene sets, and performed extensive validation and comparison against existing tools. Our tool is implemented in the existing partis annotation and clonal family inference package. We have shown, first, that the currently widespread practice of aligning expressed BCR samples against the full IMGT germline set results in both large numbers of spurious alleles and inaccurately inferred naive ancestors. Second, we showed that our method infers significantly more accurate germline sets than the existing TIgGER and IgDiscover methods in terms of both inferred gene similarity and naive ancestor inference, but of similar accuracy in terms of raw number of genes. We also showed that so far as we can determine using a wide variety of comparisons, our method's performance on real data is similar to that on simulation.

While our method has reached a level of maturity such that it provides reliable general performance, and as such is now run by default in the partis annotation and clonal family inference procedures, it has a number of weaknesses. First, as with the rest of partis, it assumes that all corrections for sequencing error have been performed before input. Second, our piecewise-linear model for the mutation accumulation plots (see Methods) is only an approximation of the real behavior. Thus, while we have designed our method with the aim of maximizing robustness against atypical repertoires, a more complex model that more closely modeled the repertoire's nonlinearities would provide better performance. Another source for improved performance would be the incorporation of per-base mutation information, i.e. splitting apart the mutation accumulation plots by A, C, G, and T. Additionally, because we do not yet set any prior on the number of germline V genes, our method will underestimate this number on smaller samples (roughly a few thousand sequences or less). Also, we have thus far only applied our method to V region genes, although the extension to D and J should be conceptually straightforward. Finally, taking advantage of the fact that rearrangement occurs only between genes on the same chromosome, as in [8, 33], would likely provide additional improvement.

A further limitation of our method is that it looks only for new alleles separated by SNPs from existing alleles, and not for those separated by insertion/deletion events. While this is not a significant limitation on human samples, the IMGT germline sets for other species are incomplete enough that, in those species, this could cause novel alleles to be misinterpreted as SHM indels. This is one respect in which the clustering-based approach taken by IgDiscover offers a significant advantage (see Supplement). For this reason we have also implemented a non-default clustering-based method which can be run in addition to the purely mutation accumulation plot-based method described here (see Manual). While we thus recommend this clustering-based method for non-human samples, its robustness, like IgDiscover's, can suffer on some highly-mutated samples, so we have left it as a non-default option pending future improvements.

Methods

Overview. The task of inferring germline genes consists largely of learning to distinguish between positions that are highly mutated as a result of SHM, and those whose highly-mutated appearance stems from the occurrence of previously unknown alleles. A few key observations allow us to extract enough information to make this distinction. First, in the absence of unknown alleles, the probability of a mutation at each position in an observed sequence is roughly proportional to the total number of mutations in that sequence (at least at the low SHM levels relevant for new-allele inference). In other words, while mutation rates differ dramatically from position to position according to, for instance, hot and cold spot motifs, each position is more likely to be mutated in sequences that have been subject to higher levels of SHM. In the presence of unknown alleles, on the other hand, sequences stemming from these unknown alleles will be mistakenly assigned to the most similar known allele, causing this approximate proportionality to be violated. If there are, say, N_{snp} SNPs separating a known and unknown allele, then there will be very few sequences from this unknown allele that appear to have fewer than N_{snp} mutations. The N_{snp} positions at which they differ, on the other hand, will almost always appear to be mutated in sequences that appear to contain N_{snp} or more total mutations. This differing apparent mutational behavior between sequences with fewer than, as compared to more than, N_{snp} mutations provides the basis for our method.

A convenient way to visualize these observations is with a type of plot introduced in [11], which we call a "mutation accumulation" plot. To make a set of these plots for one germline gene, we first collect every sequence in the sample that aligns most closely to this single known gene. We then use these sequences to make one plot for each nucleotide position as follows. The sequences are binned along the x-axis according to their total number of apparent V mutations. The y-coordinate of each bin, meanwhile, is the frequency with which that plot's nucleotide position appears to mutate among the sequences in the bin. For the full repertoire, we first group sequences based on their closest known germline gene, and then follow the procedure above for each such group. We first show example plots for three simple, hypothetical repertoires (Fig 14). While these simple repertoires, by themselves, are gross simplifications of the biological complexity in a real BCR repertoire, they contain the essential elements from which we can construct a method that performs well on real data sets.

Models and fitting. In the context of mutation accumulation plots (Fig 14), the presence of new alleles is signaled by a departure from what would be expected if all sequences had been assigned to the correct true gene. Namely, to the extent that mutations at each site accumulate in proportion to the total number of mutations in the sequence, correct assignment would result in simple linearity. For incorrect assignment, this linearity is replaced with differing behavior between the regions below and above N_{snp}. Our task, then, amounts to distinguishing between plots that can be adequately described by a one-piece linear model, and those that require a model consisting of two pieces separated by a discontinuity.

In order to distinguish these two hypotheses, we construct a model for each. The one-piece model is simply a linear fit constrained to pass through the origin. The two-piece model, meanwhile, consists of two separate linear fits, which we call the "lower" (below N_{snp}) and "upper" (above N_{snp}) fits. The lower fit is constrained to pass through the origin, while the upper fit's y-intercept must be near the average of the upper-region mutation frequencies (within 1.5 standard deviations of their mean). The junction between the two pieces must harbor a significant discontinuity in either bin value (mutation frequency) or bin total (number of sequences per bin), where significance is defined as a difference of more than 2.5 times the larger uncertainty. This two-piece model describes the presence of a new allele separated by N_{snp} SNPs from the original known gene. To give a general idea of the implementation, several examples are shown in Fig 15.

We use a ratio of error descriptors to determine whether a plot is adequately described by the one-piece fit. Define ϵ to be the sum of squared residuals divided by degrees of freedom, which in regression analysis is sometimes called the mean squared error. Good fits are characterized by values of ϵ around one, while values much greater than one indicate poor fits. Values significantly less than one generally indicate poorly-estimated uncertainties. For our purposes, then, we are interested in positions (which we call "candidate positions") for which ϵ is large for the one-piece fit (greater than 4.5) but around one (less than 1.95) for the two-piece fit.

For each N_{snp}, we construct the most plausible potential new allele by finding the N_{snp} positions that have the worst one-piece, but best two-piece, fits. We quantify this using the ratio of the two ϵ,

$$r = \frac{\epsilon_{\text{1-piece}}}{\epsilon_{\text{2-piece}}}. \qquad (1)$$

Because cases that would be better described by more complex models will have larger residuals (poor fits) for both one-piece and two-piece models, which cancel out in the ratio, this formulation provides robustness to deviations from linearity. The model for the best potential new allele consists of the N_{snp} positions that have the largest values of r.

In order to strike an appropriate balance between focusing the fit's attention on the area of the discontinuity, while taking advantage of the largest possible sample size from many surrounding bins, we perform all fits in a window of width 10 bins. This window begins at zero for small N_{snp}, while for larger N_{snp} it is symmetric around N_{snp}.

We apply several additional criteria to ensure that the candidate fits make a compelling case for a new allele. The slope at the discontinuity, i.e. the slope defined by the two points on either side, must be much larger than both the upper and lower fitted slopes (a fractional difference of more than 2.5 times). For larger N_{snp} (five or more), the slopes before and after the discontinuity must also either be consistent, or the lower slope must be the smaller of the two.

The unfortunate profusion of constant values in the preceding paragraphs deserves some examination. In general, for the sake of simplicity and interpretability, we have wherever possible minimized the number of such constants. However, practical constraints make it difficult to reduce their number further. In theory, it would be possible to construct a more complicated model that faithfully recreated all the details of the real system, which would enable a collection of simple likelihood ratio tests. However, in practice this approach is unlikely to be computationally feasible, and would likely require a much lengthier development process. Instead, we have adopted the approach of comprehensively validating a simpler model which, nevertheless, provides an adequate description of the system's real biological complexity. This method's robust performance across a wide variety of data and simulation samples during this validation (only a small fraction of which appear in this paper) gives us great confidence in its general applicability.

Comparing multiple hypotheses. The previous section outlines a procedure for identifying a single potential new allele for each individual N_{snp}. In realistic samples, however, we must treat the general case where there may be several new alleles, either with the same N_{snp}, or spread among several N_{snp}.

To do this, we first sort every candidate position within each N_{snp} by decreasing r. In order to better adjudicate between ties in the first sort, we then sort again either by decreasing y-intercept (if N_{snp} less than three) or decreasing two-piece fit ϵ. The first N_{snp} elements of this sorted list of candidate positions are then taken as a candidate allele, the next N_{snp} positions are taken as a second candidate allele, and so on, until fewer than N_{snp} remain. The second sorting step serves to group together positions with similar fit properties, and that are thus most likely to come from the same new allele. These fit properties are affected by several aspects of the new alleles, most notably their prevalence. In cases with two new alleles with the same prevalence, for example, this is not an effective means of determining which positions go with which allele; however, in real data such cases are very rare.

For each of these candidate alleles, both the smallest r among their positions, and the mean, must be greater than 2.75. The discontinuities for every pair of positions must also be compatible, defined as the difference in bin totals (number of sequences) on either side of N_{snp}, which must be closer than three times the maximum of their two uncertainties.

This procedure is repeated for each N_{snp}, resulting in a list of candidate alleles from each; these lists are then merged into a final list that is sorted by decreasing N_{snp}. We then go through this list and discard alleles that share any positions with an allele earlier in the list. This last sorting is due to the fact that it is easier for a high-N_{snp} allele to mimic a low-N_{snp} allele than the reverse.

Approximations and pre-filters. The procedure described above would work well in principle, but would require a computationally prohibitive number of fits. As a rough estimate, taking 50 initial, known alleles in a sample, each with 300 positions, looking up through N_{snp} equals eight, and with both the one-piece and two-piece fits, we would need 360,000 individual linear fits. To be useful, however, it must run as part of the overall partis annotation, which takes only minutes on samples with tens of thousands of sequences. Luckily, the overwhelming majority of these fits can be avoided by ignoring uninteresting positions using a number of approximation procedures. The cumulative effect of the following approximations and filters is that a typical run requires of order 100 fits, with no appreciable decrease in precision or sensitivity. This results in a method that does not add significant run time to an existing partis run.

The first step is to ignore positions for which there would not be enough statistical power to have any sensitivity to new alleles. We thus skip positions with fewer than 150 total observed sequences, summed over bins. Positions with fewer than 30 observed mutations, also summed over bins, are similarly skipped.

For each N_{snp}, we also ignore positions that do not have at least eight observed mutations in the $N_{\text{snp}}{}^{\text{th}}$ bin. This bin is of primary importance, because it is the means by which we determine that this N_{snp} is the correct one, rather than those slightly larger or smaller. If this bin is truly signaling a new allele, then it must contain a significant number of mutated sequences.

For several of the subsequent steps, we use an approximate fitting procedure to arrive at a slope, intercept, and associated uncertainties. While less accurate, and more heuristic, than the least-squares fits that are used elsewhere, it is also much faster. We begin by calculating the two-point slope between each pair of adjacent points. If there are only two points in total, this is supplemented by a "synthetic" slope between the first point shifted up, and the second shifted down, by their respective uncertainties. The approximate slope is then calculated as the mean of these pairwise slopes, with its uncertainty the associated standard error. We arrive at the approximate y-intercept with a similar procedure, except that the pairwise slope is replaced by the pairwise y-intercept, which uses the previously-calculated pairwise slope.

For smaller N_{snp} (three or less), we also require that the approximate upper-region y-intercept fit bounds do not include zero. If they do include zero, there will not be a significant difference between the one- and two-piece fits. As an additional, and more stringent, test that the upper-region y-intercept for these smaller N_{snp} is well above zero, we require that the approximate fit's y-intercept is also greater than zero.

For N_{snp} equals two, we also require that the bin immediately before the $N_{\text{snp}}{}^{\text{th}}$ bin be outside of the upper-region y-intercept fit bounds.

And finally, for larger N_{snp} (five or greater), the approximate lower fit's slope must be less than that of the approximate upper fit, in cases in which they are inconsistent.

Excluded bases on 5′ and 3′ ends. We are typically analyzing only partial V sequences, which leads to additional complications. On the 5′ end, the method must account for samples in which the read does not extend through the entire V gene. On the 3′ end, meanwhile, VDJ rearrangement itself deletes some number of bases.

The presence of incomplete V sequences clearly reduces our sensitivity to new alleles simple by reducing the sample size for positions at each end. A more serious problem, however, is that differing lengths cause sequences to be assigned to incorrect bins, since their apparent number of mutations is different than their true number. In order to avoid this problem, we ensure that all analyzed sequences begin and end at the same aligned germline bases. To accomplish this, for each end (5′ and 3′), we find the deletion length such that only a small fraction (f, by default 0.01) of sequences have a longer deletion length. We then exclude the fraction f of sequences that have longer deletions. Among the remaining sequences, we then exclude from the analysis the positions that fall within these deletion lengths. For example, if 99% of sequences have 3′ deletions of four or fewer bases, then we would discard sequences with more than four 3′-deleted bases, and would not use that gene's four most-3′ positions in the fitting procedure. Note that on the 3′ side of V, this exclusion procedure is especially important because the final few germline-aligned positions next to any non-templated insertion always have very poorly-measured mutation frequencies.

Collapsing clones. Our method requires that we consider only independent mutation events, excluding any mutations that share a common ancestry. In order to satisfy this requirement we attempt to select from each clone the largest possible set of sequences without shared mutations. In doing this, we give preference to relatively unmutated sequences, since most new alleles are separated by only a few SNPs from known alleles. Specifically, we sort the sequences from each clonal family in order of increasing apparent V mutation. We then traverse this list, selecting each sequence that does not share any mutations with a previously-selected sequence. As shown in [40] it would be straightforward to use the full partis method to separate the sequences into clonal families. However, for the purposes of ensuring independent mutations, there is little benefit to having precisely accurate clusters, since a slightly inaccurate clustering only results in slightly inaccurate uncertainties in the fits, and uncertainties on uncertainties are in practice never large enough to impact the analysis. For the sake of speed, then, we simply cluster using inferred naive sequences, i.e. every sequence that is inferred to have the same naive sequence is clustered together. This has the additional benefit of becoming more conservative as the sample size becomes large – in other words it tends to over-cluster more as the space of potential naive rearrangements fills up, and nearby rearrangement events have very similar naive sequences. This has the effect of sacrificing some sensitivity in order to ensure that mutations are actually independent.

Initial removal of less-likely alleles. Some care is necessary when constructing each sample's initial set of known genes. We find the performance of our new-allele inference to be robust enough that the best approach is to first choose a minimal number of genes whose presence in the sample is supported by very strong evidence. We then apply the new-allele inference framework in order to reinstate alleles for which the evidence was less overwhelming, along with any novel alleles.

In order to construct each sample's minimal initial gene set, we first partition the complete set of IMGT [1] genes into groups within which SHM can easily cause confusion, and then retain only the most common gene in each group. Note that this partitioning cannot be accomplished using only the IMGT names – there are many cases of allelic variants that differ by so many SNPs that confusion is very unlikely, as well as alleles of separate genes that differ by only a single SNP. We construct these groups by single-linkage clustering such that genes with the same conserved cysteine position, and separated by fewer than eight SNPs, are grouped together. In order to ensure that we can re-infer all of the genes within each group, this number corresponds to the maximum number of SNPs for new-allele inference.

We also discard alleles that appear to occur at extremely low frequencies, by default less than one part in 2000.

Template allele removal. The procedure outlined thus far can yield a confident judgment on whether there exists a previously-unknown allele separated from some known, “template” allele. We must also, however, distinguish between cases where this template allele is also present in the sample, and cases where it is not (and was simply the closest known allele). In order to do this we observe that in the plots in Fig 15 the y-intercept of the upper (post-N_{snp}) fit is determined largely by the prevalence of the new allele. For N_{snp} near one, the y-value is very close to the actual allele prevalence, while for larger N_{snp} the relationship is more complex. When the new allele’s prevalence is 1, i.e. the template allele is not present in the sample, however, the fitted post-N_{snp} y-value is also very close to 1. The only deviation is a slight decreasing slope from reversion to germline at higher mutation levels. We thus remove template genes from the germline set when the upper fits for each position have y-intercept 1.1 ± 0.12 and slope -0.01 ± 0.015.

Adding a new allele. Once we have decided that there is sufficient evidence for a new allele separated by N_{snp} SNPs from an existing allele, there remain several additional considerations.

First, we must determine its original germline sequence. We begin by restricting ourselves to sequences assigned to the $N_{\mathrm{snp}}{}^{\mathrm{th}}$ bin, i.e. which contain N_{snp} apparent mutations. This restriction is important, because unmutated sequences stemming from the new allele are assigned to this bin. It thus minimizes the confusion caused by mutated sequences derived from the existing allele, as well as from any additional new alleles. For each of the N_{snp} positions where this allele differs from the template allele, we then choose as the new allele’s germline nucleotide the most commonly-observed non-template nucleotide at that position.

If the newly-identified allele was present in the original, full germline set, then we add it with its original name; otherwise we add it with a provisional name derived from the template gene. Because of the unavoidable ambiguity created by 3′ exonuclease deletion (and short reads), in order to be considered equivalent we require only that two alleles are identical after applying the 5′ and 3′ exclusions described above. If, for instance, we infer a new allele that differs by several SNPs from some template gene, and there is an existing allele in the original set that is identical to this new allele except for an extra base to the 3′ of the cysteine, we assume that the newly-inferred and existing alleles are in fact the same. More generally, we note that any new-allele inference framework that uses expressed data will suffer from a large uncertainty as to the precise number and identity of a V gene’s most-3′ few germline bases. In order to resolve this uncertainty we must perform germline sequencing.

Simulation details. The simulated samples used for validation were made with the same basic framework described in [4]. In addition to the details described there, we have added options to control various aspects of a sample's germline set. All simulation options are described in detail in the manual (`https://git.io/vFKok`).

Most basically, we have added the ability to insert into a germline set new alleles that are separated from existing alleles by both point and insertion/deletion mutations. The number of each type of mutation, and their properties, are specified with command line options. Each mutation occurs either at a specified position in the allele's sequence, or at a random position within specified bounds (for instance, within vs outside of the CDR3). These options allowed the creation of the simplified sparse gene repertoire samples in the Results.

These sparse repertoires are built around a single known germline gene. We then add either one or two novel alleles, separated by SNPs at uniformly-selected random positions, from this existing germline gene.

In order to generate a germline set for the full repertoire samples, for each region we first choose some number of genes from the IMGT set, and then some number of alleles for each of these genes. The mean number of alleles per gene is specified on the command line, then for each gene we choose a number of alleles from $\{1, 2\}$ with weights such as to (on average) arrive at the specified mean over all genes. For both the "low-SHM" and "high-SHM" full-repertoire samples, this procedure was followed with 42, 22, and 6 genes (V, D, and J regions), with a mean alleles per gene of 1.33, 1.1, and 1 (V, D, and J). This is concordant with the references in Results above, in particular [7], which reported a mean over 12 individuals of 40.2 homozygous, 8.6 two-allele heterozygous, and 1.1 three-allele heterozygous V genes, for an overall mean alleles per gene of 1.2. Six novel alleles were then added, separated by 1, 1, 2, 3, 5, and 6 point mutations (at uniformly-selected random positions) from an existing allele.

We choose each gene's relative prevalence counts from a uniform random distribution with bounds $[1, 1/f_{\mathrm{min}}]$, where f_{min} is the minimum desired prevalence ratio between any pair of genes in the repertoire. This ensures that the prevalence ratio for every pair of genes in the repertoire is in $[f_{\mathrm{min}}, 1]$, for f_{min} equals 0.15 ("low-SHM" samples) or 0.05 ("high-SHM" samples). While this is roughly compatible with the variation in expression levels typically reported in real data, we emphasize that most previous studies (including our own [4]) have aligned against the full IMGT set, and as such their reported expression levels for less-common genes are probably meaningless (Table 1, Fig 4).

The SHM distributions in the full-repertoire samples were chosen to be representative of typical IgM-specific data ("low-SHM", mean value 0.02) or typical unsorted samples ("high-SHM", mean value 0.06) (compare to mean values in Fig S1).

Finally, we must decide on the clonal family structure of each sample. Real repertoires vary widely in both their clonality and lineage structure. However, we have shown in Figs 1 and 2 that our clonal family collapse is an effective-enough approximation that changes in clonality and lineage structure are equivalent to changes in sample size, and thus only affect sensitivity. In order to maximize the variety of interesting variables over which we can perform validation, we thus simulate the full repertoire samples with singleton clonal families.

Phylogenetic comparison plots. In order to make the phylogenetic gene set comparison plots (Figs 4, 5, 6, 7 8, 9, 10, S2, S3 11, 12, 13) we begin by aligning all the genes that we want to compare using MUSCLE [41] (v3.8.31 with default parameters). We then use RAxML [42] (v8.2.10, with the GTR model) to create a tree for these genes. In order to allow easy visual comparison of the entire germline gene set in one plot, while also allowing comparison within each gene family (IMGT definition, e.g. IGHV3), we then collapse to length zero each branch that joins two different gene families. If the reader would like to compare combinations of germline sets that are not shown in this paper, all true and inferred germline sets, for simulation and data, are available at `https://zenodo.org/record/1037464#.WfISc3BrwUE`, and the command line script used to make these plots is available at `https://git.io/vFo2B`.

Acknowledgements

The authors would like to thank Kristian Davidsen, Laura Noges, and Peter Ralph for helpful discussions. This research was supported by NIH grants U19-AI117891, R01-GM113246, and R01-AI12096. The research of Frederick Matsen was supported in part by a Faculty Scholar grant from the Howard Hughes Medical Institute and the Simons Foundation.

References

1. Lefranc MP, Giudicelli V, Ginestoux C, Jabado-Michaloud J, Folch G, Bellahcene F, et al. IMGT, the international ImMunoGeneTics information system. Nucleic Acids Res. 2009 Jan;37(suppl 1):D1006–D1012. Available from: `http://nar.oxfordjournals.org/content/37/suppl_1/D1006.abstract`.
2. Ye J, Ma N, Madden TL, Ostell JM. IgBLAST: an immunoglobulin variable domain sequence analysis tool. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W34–40. Available from: `http://dx.doi.org/10.1093/nar/gkt382`.
3. Brochet X, Lefranc MP, Giudicelli V. IMGT/V-QUEST: the highly customized and integrated system for IG and TR standardized V-J and V-D-J sequence analysis. Nucleic Acids Res. 2008 Jul;36(Web Server issue):W503–8.
4. Ralph DK, Matsen FA IV. Consistency of VDJ Rearrangement and Substitution Parameters Enables Accurate B Cell Receptor Sequence Annotation. PLOS Comput Biol. 2016 Jan;12(1):e1004409. Available from: `http://dx.doi.org/10.1371/journal.pcbi.1004409`.
5. Bolotin DA, Poslavsky S, Mitrophanov I, Shugay M, Mamedov IZ, Putintseva EV, et al. MiXCR: software for comprehensive adaptive immunity profiling. Nat Methods. 2015 29 Apr;12(5):380–381.
6. Wang Y, Jackson KJ, Gäeta B, Pomat W, Siba P, Sewell WA, et al. Genomic screening by 454 pyrosequencing identifies a new human IGHV gene and sixteen other new IGHV allelic variants. Immunogenetics. 2011 May;63(5):259–265.
7. Boyd SD, Gaëta BA, Jackson KJ, Fire AZ, Marshall EL, Merker JD, et al. Individual variation in the germline Ig gene repertoire inferred from variable region gene rearrangements. J Immunol. 2010 Jun;184(12):6986–6992.
8. Kidd MJ, Chen Z, Wang Y, Jackson KJ, Zhang L, Boyd SD, 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 1 Feb;188(3):1333–1340.
9. Watson CT, Steinberg KM, Huddleston J, Warren RL, Malig M, Schein 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 4 Apr;92(4):530–546.

10. Watson CT, Steinberg KM, Graves TA, Warren RL, Malig M, Schein J, 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 23 Oct;Available from: http://dx.doi.org/10.1038/gene.2014.56.
11. 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. Proc Natl Acad Sci U S A. 2015 9 Feb;.
12. Corcoran MM, Phad GE, Vázquez Bernat N, Stahl-Hennig C, Sumida N, Persson MAA, et al. Production of individualized V gene databases reveals high levels of immunoglobulin genetic diversity. Nat Commun. 2016 Dec;7:13642.
13. Scheepers C, Shrestha RK, Lambson BE, Jackson KJL, Wright IA, Naicker D, et al. Ability to develop broadly neutralizing HIV-1 antibodies is not restricted by the germline Ig gene repertoire. J Immunol. 2015 May;194(9):4371–4378.
14. Lee CEH, Gaëta B, Malming HR, Bain ME, Sewell WA, Collins AM. Reconsidering the human immunoglobulin heavy-chain locus:. Immunogenetics. 2006;57(12):917–925.
15. Lee CEH, Jackson KJL, Sewell WA, Collins AM. Use of IGHJ and IGHD gene mutations in analysis of immunoglobulin sequences for the prognosis of chronic lymphocytic leukemia. Leuk Res. 2007 Sep;31(9):1247–1252.
16. Wang Y, Jackson KJL, Sewell WA, Collins AM. Many human immunoglobulin heavy-chain IGHV gene polymorphisms have been reported in error. Immunol Cell Biol. 2008 Feb;86(2):111–115.
17. Collins AM, Wang Y, Roskin KM, Marquis CP, Jackson KJL. The mouse antibody heavy chain repertoire is germline-focused and highly variable between inbred strains. Philos Trans R Soc Lond B Biol Sci. 2015 Sep;370(1676). Available from: http://dx.doi.org/10.1098/rstb.2014.0236.
18. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science. 2001 Feb;291(5507):1304–1351. Available from: http://dx.doi.org/10.1126/science.1058040.
19. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, et al. Initial sequencing and comparative analysis of the mouse genome. Nature. 2002 Dec;420(6915):520–562.
20. 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, et al. A map of human genome variation from population-scale sequencing. Nature. 2010 Oct;467(7319):1061–1073. Available from: http://dx.doi.org/10.1038/nature09534.
21. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet. 2011 Nov;13(1):36–46.
22. Chimge NO, Pramanik S, Hu G, Lin Y, Gao R, Shen L, et al. Determination of gene organization in the human IGHV region on single chromosomes. Genes Immun. 2005 May;6(3):186–193.
23. Luo S, Yu JA, Li H, Song YS. Worldwide genetic variation of the IGHV and TRBV immune receptor gene families in humans. bioRxiv. 2017 Jun;p. 155440. Available from: http://dx.doi.org/10.1101/155440.
24. Yu Y, Ceredig R, Seoighe C. A Database of Human Immune Receptor Alleles Recovered from Population Sequencing Data. J Immunol. 2017 Jan;Available from: http://dx.doi.org/10.4049/jimmunol.1601710.
25. Watson CT, Matsen IV FA, Jackson KJL, Bashir A, Smith ML, Glanville J, et al. Comment on "A Database of Human Immune Receptor Alleles Recovered from Population Sequencing Data". The Journal of Immunology. 2017 May;198(9):3371–3373. Available from: http://www.jimmunol.org/content/198/9/3371?etoc.
26. Avnir Y, Watson CT, Glanville J, Peterson EC, Tallarico AS, Bennett AS, et al. *IGHV1-69* polymorphism modulates anti-influenza antibody repertoires, correlates with IGHV utilization shifts and varies by ethnicity. Sci Rep. 2016 Feb;6:20842.
27. Liu L, Lucas AH. IGH V3-23*01 and its allele V3-23*03 differ in their capacity to form the canonical human antibody combining site specific for the capsular polysaccharide of Haemophilus influenzae type b. Immunogenetics. 2003 Aug;55(5):336–338.
28. Rogozin IB, Kolchanov NA. Somatic hypermutagenesis in immunoglobulin genes. II. Influence of neighbouring base sequences on mutagenesis. Biochim Biophys Acta. 1992 Nov;1171(1):11–18. Available from: http://www.ncbi.nlm.nih.gov/pubmed/1420357.
29. Cui A, Di Niro R, Vander Heiden JA, Briggs AW, Adams K, Gilbert T, et al. A Model of Somatic Hypermutation Targeting in Mice Based on High-Throughput Ig Sequencing Data. J Immunol. 2016 Jan;197(9):3566–3574.

30. Yaari G, Uduman M, Kleinstein SH. Quantifying selection in high-throughput Immunoglobulin sequencing data sets. Nucleic Acids Res. 2012 May;40(17):e134. Available from: `http://dx.doi.org/10.1093/nar/gks457`.
31. McCoy CO, Bedford T, Minin VN, Bradley P, Robins H, Matsen IV FA. Quantifying evolutionary constraints on B-cell affinity maturation. Philos Trans R Soc Lond B Biol Sci. 2015 5 Sep;370(1676). Available from: `http://dx.doi.org/10.1098/rstb.2014.0244`.
32. Matsuda F, Ishii K, Bourvagnet P, Kuma Ki, Hayashida H, Miyata T, et al. The complete nucleotide sequence of the human immunoglobulin heavy chain variable region locus. J Exp Med. 1998 Dec;188(11):2151–2162.
33. 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 5 Sep;370(1676). Available from: `http://dx.doi.org/10.1098/rstb.2014.0243`.
34. Vander Heiden JA, Stathopoulos P, Zhou JQ, Chen L, Gilbert TJ, Bolen CR, et al. Dysregulation of B Cell Repertoire Formation in Myasthenia Gravis Patients Revealed through Deep Sequencing. J Immunol. 2017 13 Jan;.
35. Laserson U, Vigneault F, Gadala-Maria D, Yaari G, Uduman M, Vander Heiden JA, et al. High-resolution antibody dynamics of vaccine-induced immune responses. Proc Natl Acad Sci U S A. 2014 17 Mar;Available from: `http://dx.doi.org/10.1073/pnas.1323862111`.
36. Sheng Z, Schramm CA, Kong R, Program NCS, Mullikin JC, Mascola JR, et al. Gene-Specific Substitution Profiles Describe the Types and Frequencies of Amino Acid Changes during Antibody Somatic Hypermutation. Frontiers in immunology. 2017 May;8:537.
37. Williams K, Noges L, et al. Manuscript in preparation;.
38. Hagen O, Hartmann K, Steel M, Stadler T. Age-dependent speciation can explain the shape of empirical phylogenies. Syst Biol. 2015 May;64(3):432–440.
39. Lefranc MP, Lefranc G. The Immunoglobulin FactsBook. Cambridge, MA: Academic Press; 2001.
40. Ralph DK, Matsen IV FA. Likelihood-Based Inference of B Cell Clonal Families. PLOS Comput Biol. 2016 Oct;12(10):e1005086.
41. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004 Mar;32(5):1792–1797.
42. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014 May;30(9):1312–1313.

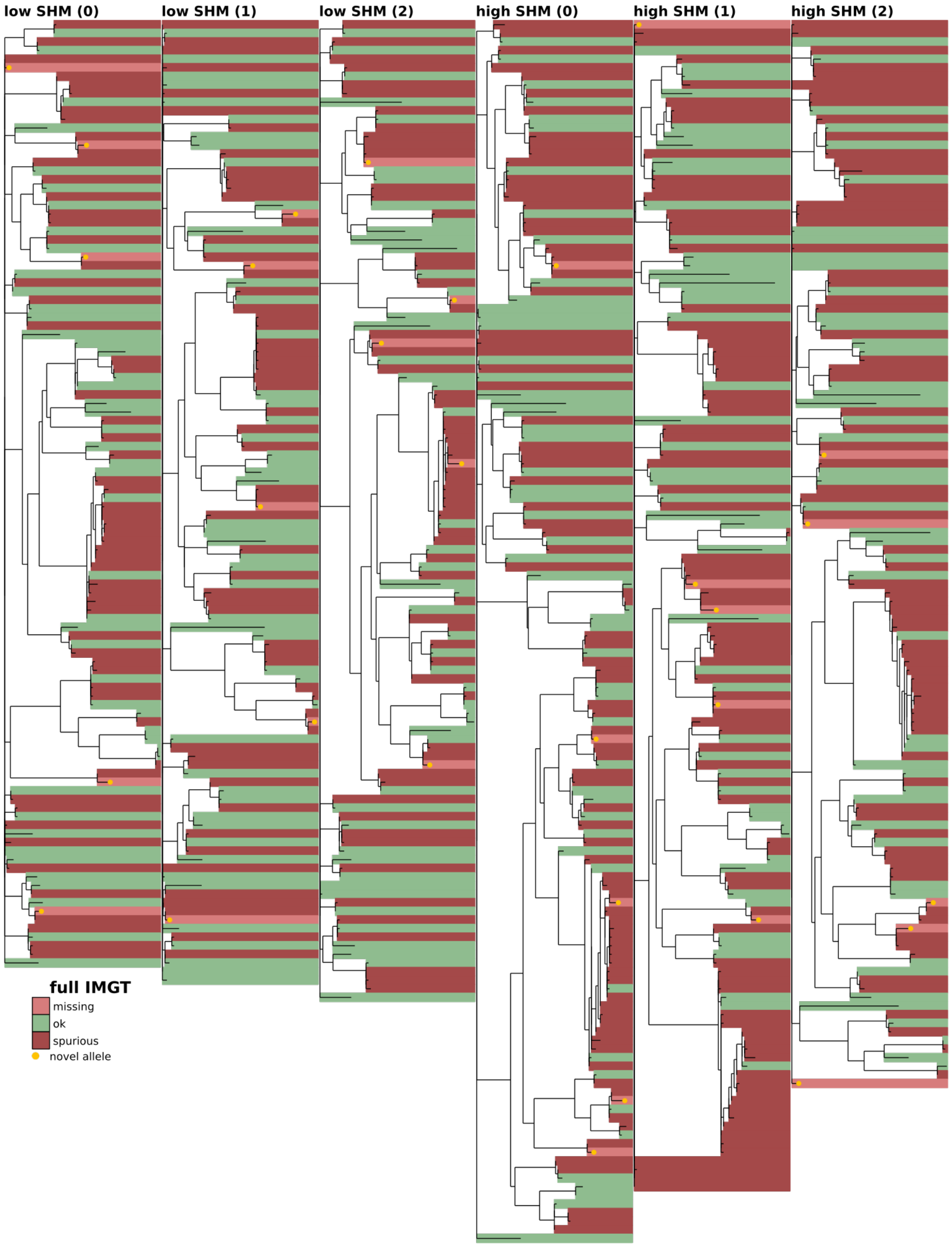

FIGURE 4. **Full-repertoire germline set accuracy for the currently widespread method of aligning every sequence to its closest match in the full IMGT V gene set**. The phylogenetic tree is constructed with a leaf for each germline gene in either the true or inferred germline sets (see Methods). Branch lengths connecting different V gene families are set to zero. Leaves are colored according to the similarity of the true and inferred germline sets, with shared genes in green and unshared in red, the latter broken into missing (light red) and spurious (dark red). Novel alleles (not in the IMGT database, whether from the true simulated set or spuriously inferred) are highlighted in gold. Shown on the first three replicates (0-2) of both the low-SHM (left), and high-SHM (right) full-repertoire simulation samples (see text).

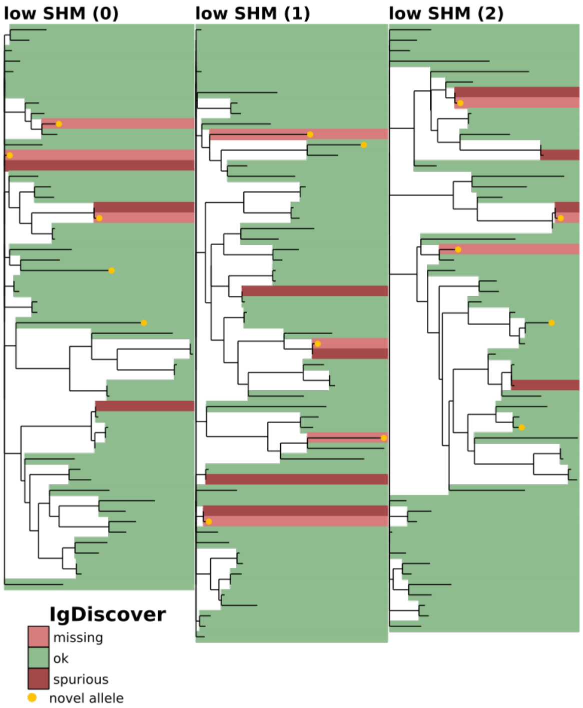

FIGURE 5. **Full-repertoire germline set accuracy for IgDiscover** (explanation in Fig 4). Shown on the first three replicates (0-2) of the low-SHM full-repertoire simulation samples. The high-SHM samples are not shown, since IgDiscover is designed only for low-SHM IgM samples (see text).

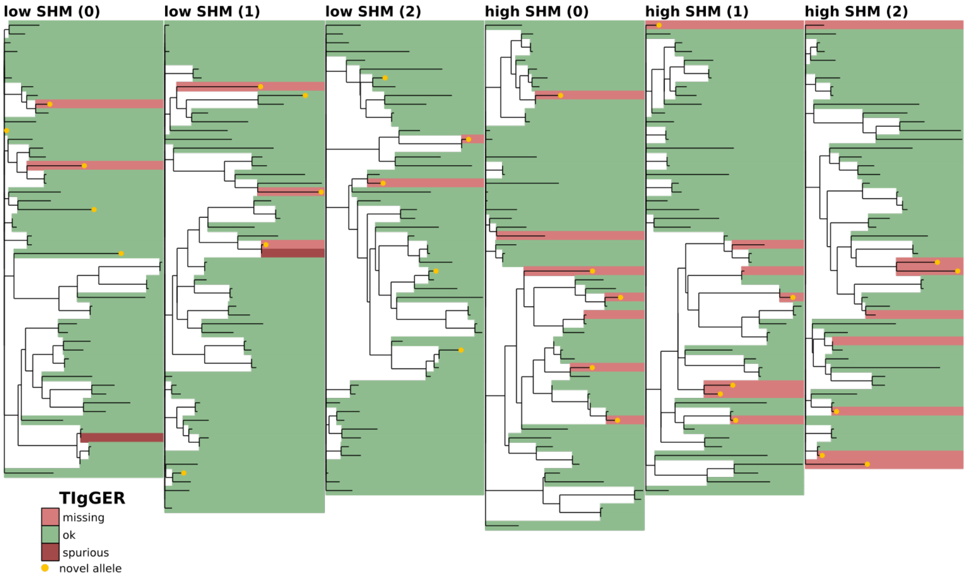

FIGURE 6. **Full-repertoire germline set accuracy for TIgGER** (explanation in Fig 4).

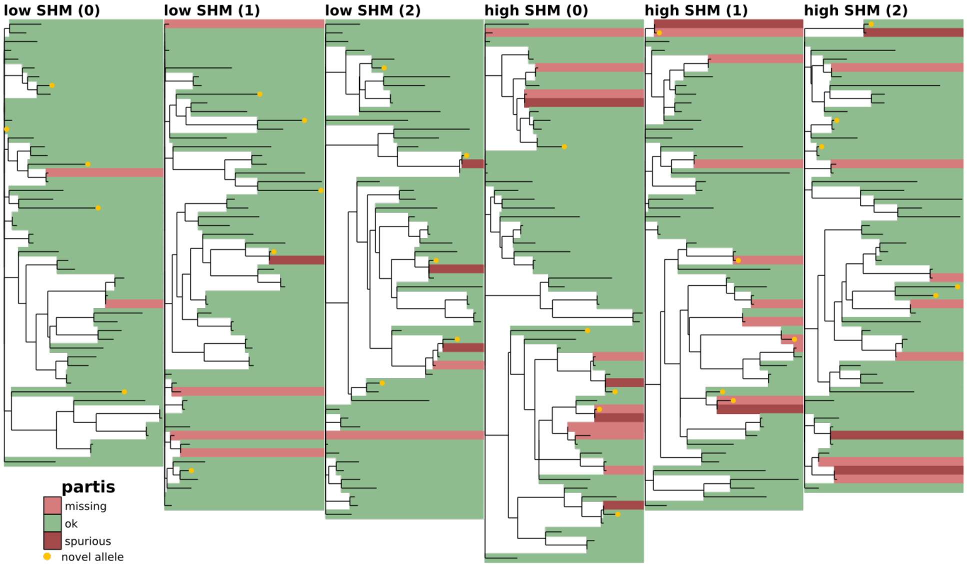

FIGURE 7. **Full-repertoire germline set accuracy for partis** (explanation in Fig 4).

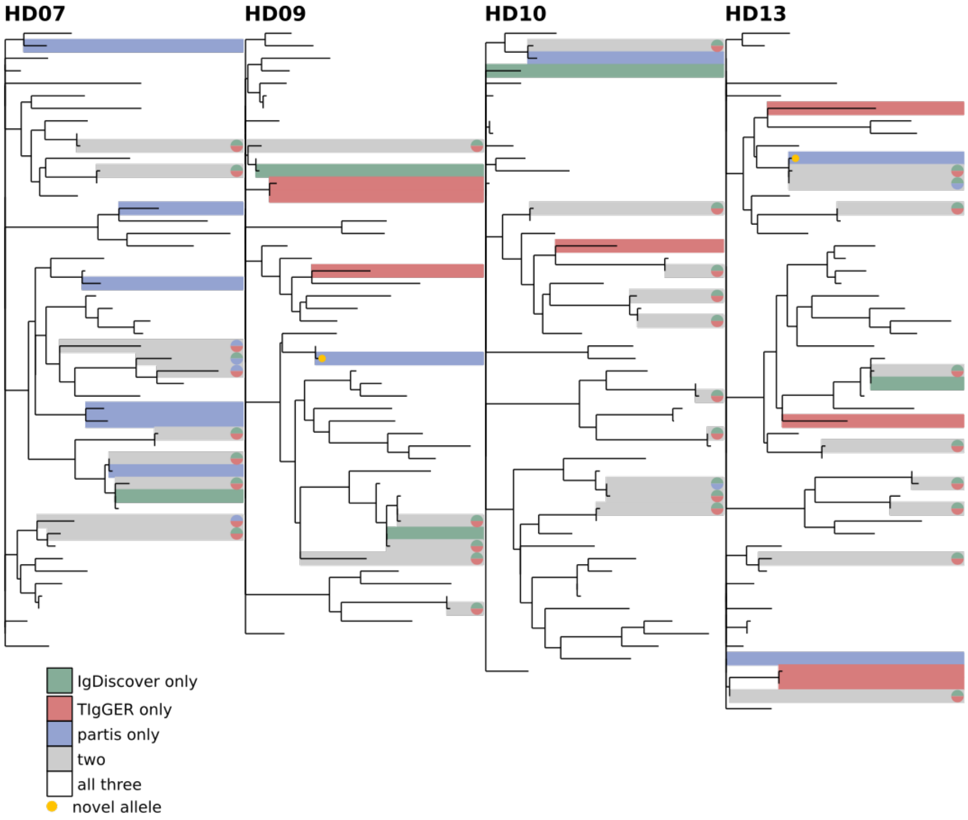

FIGURE 8. **Comparison of all three inference methods on the healthy donor samples from [34]** (other subjects shown in Figs S2 and S3). The phylogenetic tree is constructed with a leaf for each germline gene that was inferred by any of the methods. Branch lengths connecting different V gene families are set to zero. Leaves are colored according to how many methods inferred the corresponding gene: one (green, red, blue), two (grey), or all three (white).

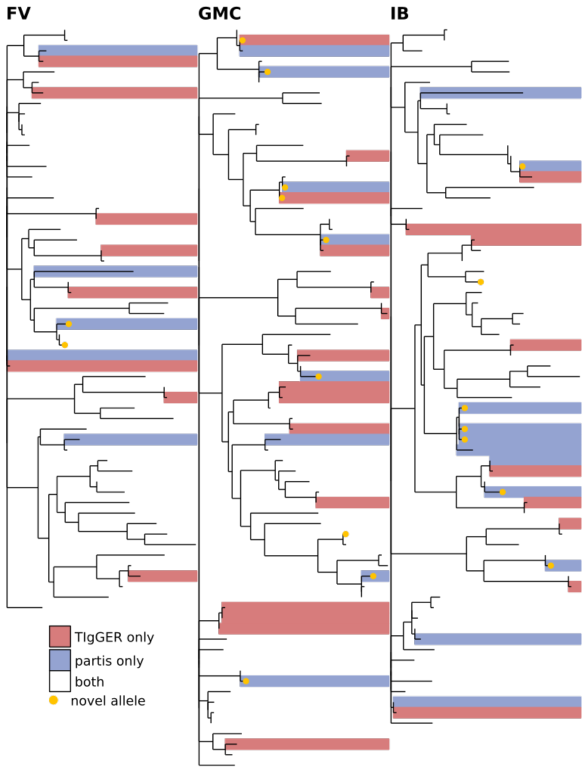

FIGURE 9. **Comparison of germline sets inferred by partis and TIgGER for subjects FV, GMC, and IB from [35]**, with all ten time points merged for each subject. The phylogenetic tree is constructed with a leaf for each germline gene that was inferred by either of the two methods. Branch lengths connecting different V gene families are set to zero. Leaves are colored according to how many methods inferred the corresponding gene: either one (red, blue) or both (white). Since this data is not IgM specific, IgDiscover is not shown. Includes the three time points in Fig 11, plus seven more, for each subject.

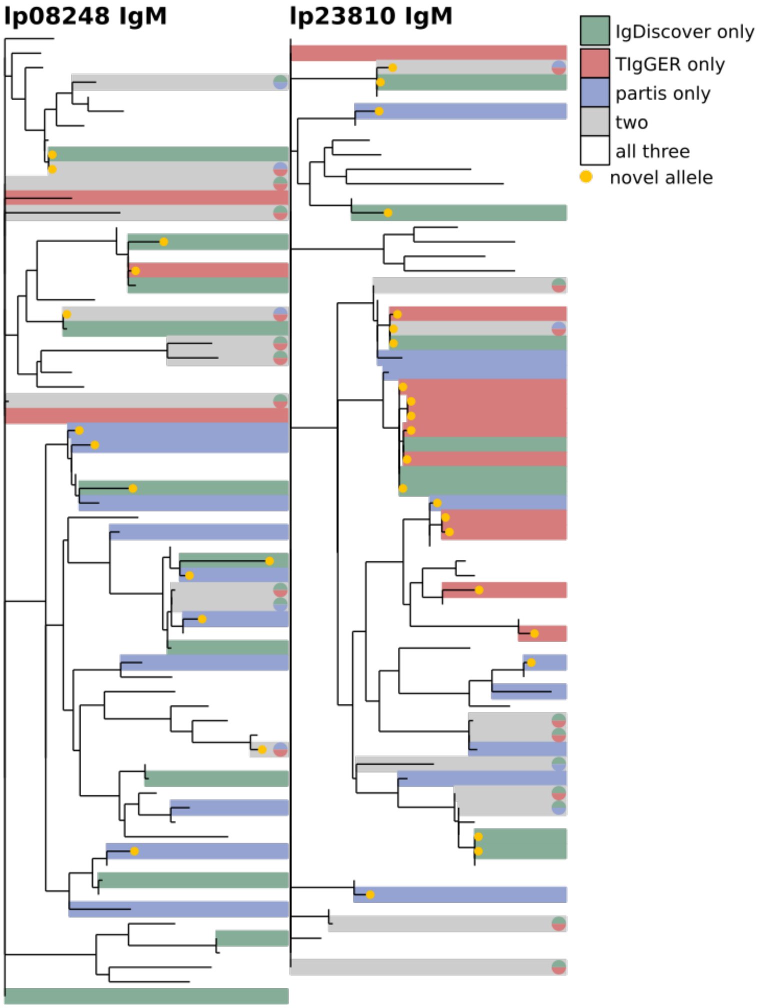

FIGURE 10. **Comparison of the three methods on IgM data** from subjects lp08248 (left) and lp23810 (right) from [36]. The phylogenetic tree is constructed with a leaf for each germline gene that was inferred by any of the methods. Branch lengths connecting different V gene families are set to zero. Leaves are colored according to how many methods inferred the corresponding gene: one (green, red, blue), two (grey), or all three (white). See Fig 13 for other results for these subjects.

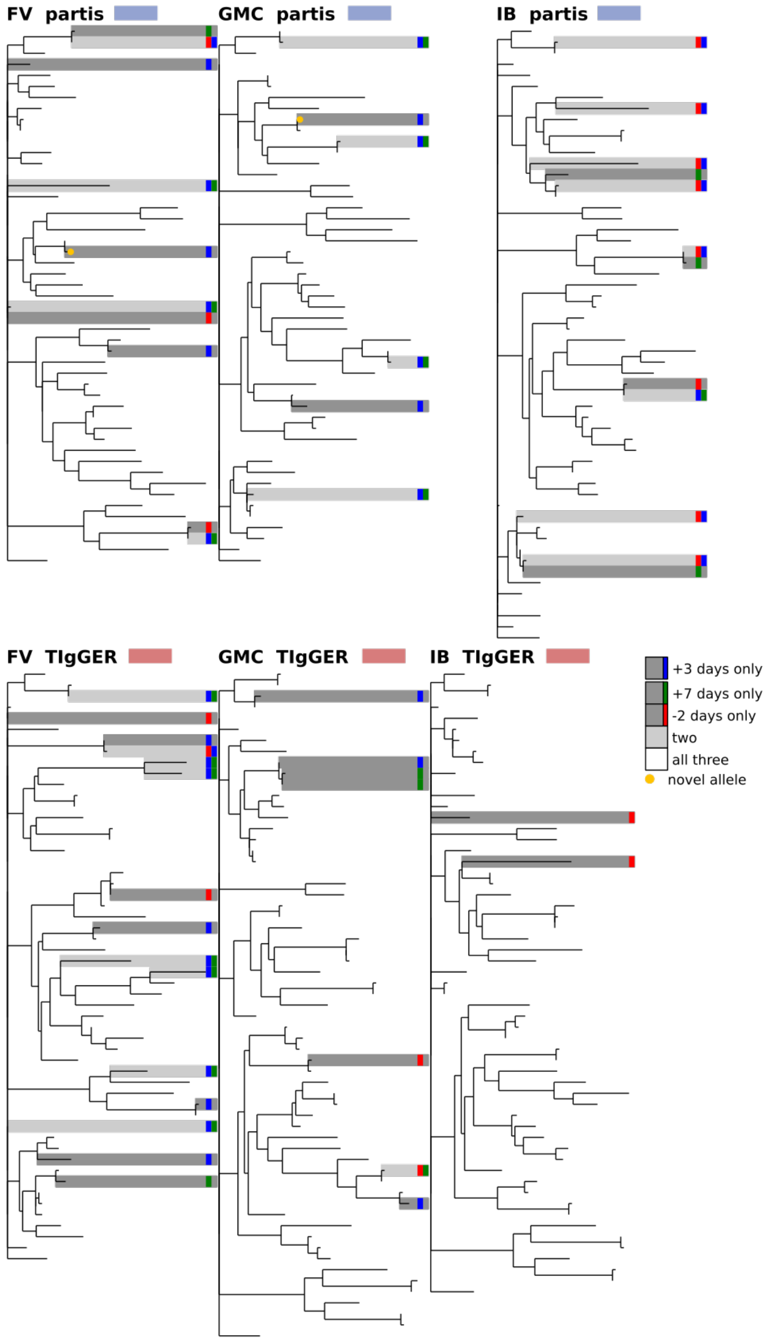

FIGURE 11. **Comparison of inferred germline sets for samples taken at different time points for subjects FV, GMC, and IB from [35].** Shown for three (of ten total) time points surrounding influenza vaccination: two days before, three days after, and seven days after; for partis (top) and TIgGER (bottom). The phylogenetic tree is constructed with a leaf for each germline gene that was inferred at any of the three time points. Branch lengths connecting different V gene families are set to zero. Leaves are colored according to the number of time points at which the corresponding gene was inferred: one (dark grey), two (light grey), or all three (white). Since this data is not IgM specific, IgDiscover is not shown. See Fig 9 for other results for these subjects.

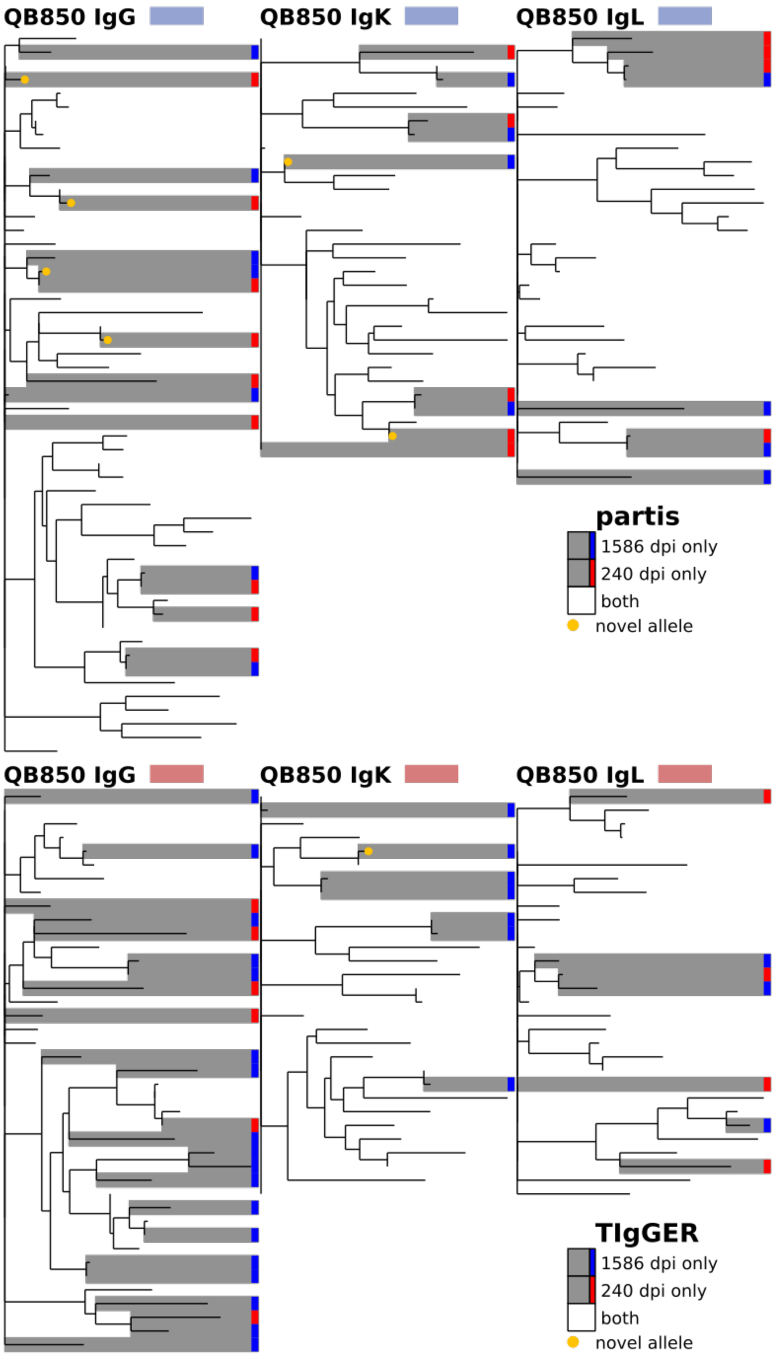

FIGURE 12. **Comparison of inferred germline sets for samples taken at two time points from the same subject** for IGH (left), IGK (middle), and IGL (right) for partis (top) and TIgGER (bottom). Time points are three years apart in the HIV-superinfected subject QB850 from [37]. The phylogenetic tree is constructed with a leaf for each germline gene that was inferred at either of the two time points. Branch lengths connecting different V gene families are set to zero. Leaves are colored according to the number of time points at which the corresponding gene was inferred: either one (grey) or both (white). Since this data is not IgM specific, IgDiscover is not shown.

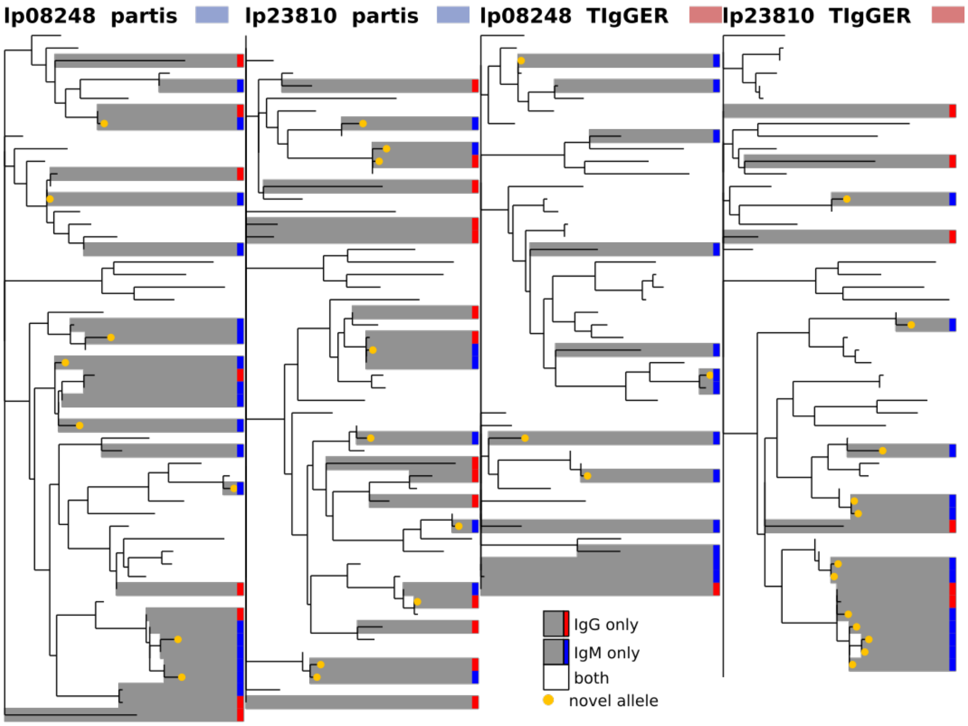

FIGURE 13. **Comparison of inferred germline sets for IgM vs IgG data** from subjects lp08248 and lp23810 from [36] for partis (left) and TIgGER (right). The phylogenetic tree is constructed with a leaf for each germline gene that was inferred for either of the two isotypes. Branch lengths connecting different V gene families are set to zero. Leaves are colored according to the number of isotype-specific samples for which the corresponding gene was inferred: either one (grey) or both (white). See Fig 10 for other results for these subjects.

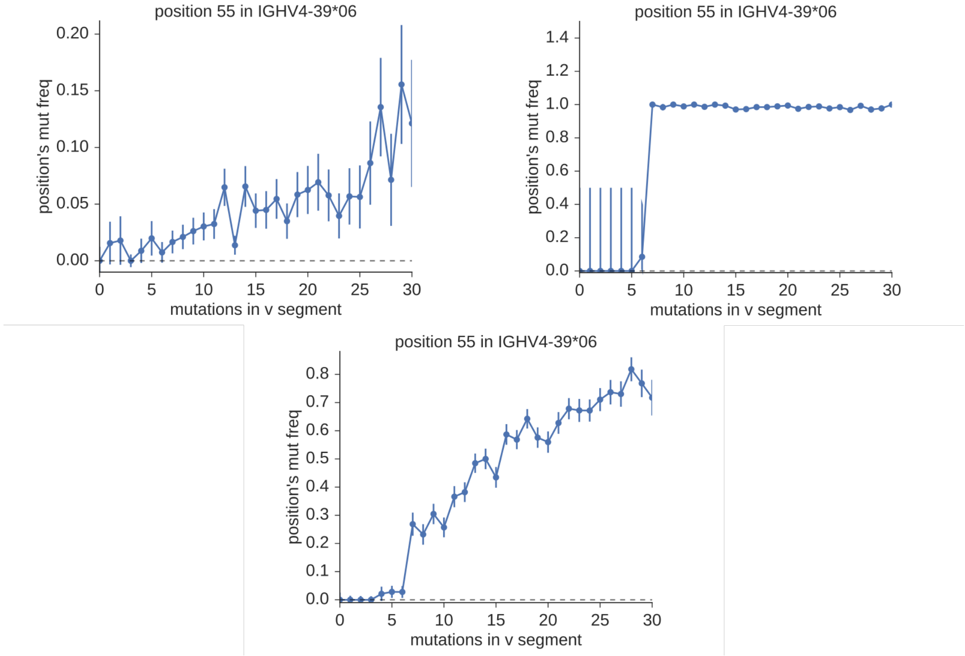

FIGURE 14. Mutation accumulation plots showing the relationship between the mutation probability at position 55 across all sequences aligning closest to IGHV4-38*06 (y-axis), and the number of mutations in the entire observed V sequence (x-axis) for three simple, hypothetical BCR repertoires. In the top row are two repertoires that consist of a single allele: where this allele is known (left), and where it is unknown, but separated by seven SNPs from a known allele (right). In a more typical case, given the relative completeness of the standard germline sets, we would observe a mixture of sequences from the known and unknown alleles (bottom). This is equivalent to the (shifted) superposition of the two plots in the top row.

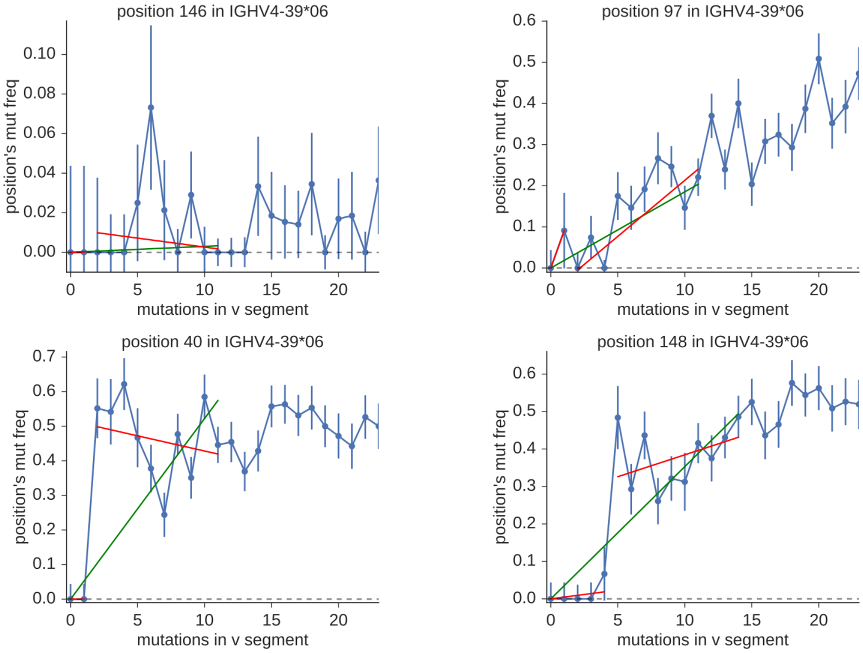

FIGURE 15. **Example one-piece (green) and two-piece (red) fits for positions without (top row) and with (bottom row) evidence for new alleles.** The left and right plots in the top row show the difference between positions with low and high mutability (cold and hot spots). The bottom row shows a position with evidence for a new allele with N_{snp} equal to two (left) and a similar plot for N_{snp} equals five (right). Note that both one-piece and two-piece models fit well in the top row, whereas in the bottom row only the two-piece model provides an adequate fit.

SUPPLEMENTARY INFORMATION

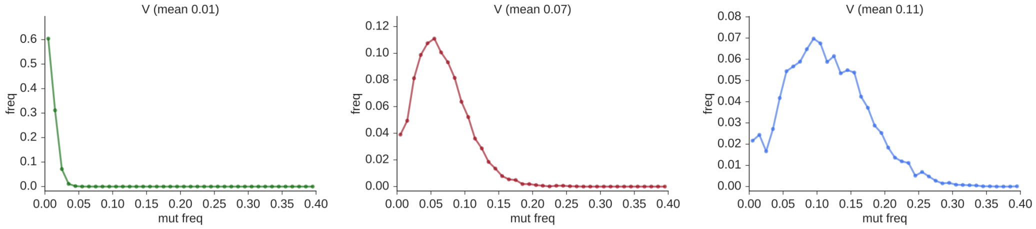

FIGURE S1. **V region mutation frequency distributions** from the low (left), typical (middle), and high (right) mutation samples used to make the top left panels of Figures 1 and 2. In the context of these distributions, the full-repertoire samples (Figs 4, 5, 6, 7) correspond to a mean value of 0.02 ("low-SHM" samples) and 0.06 ("high-SHM" samples).

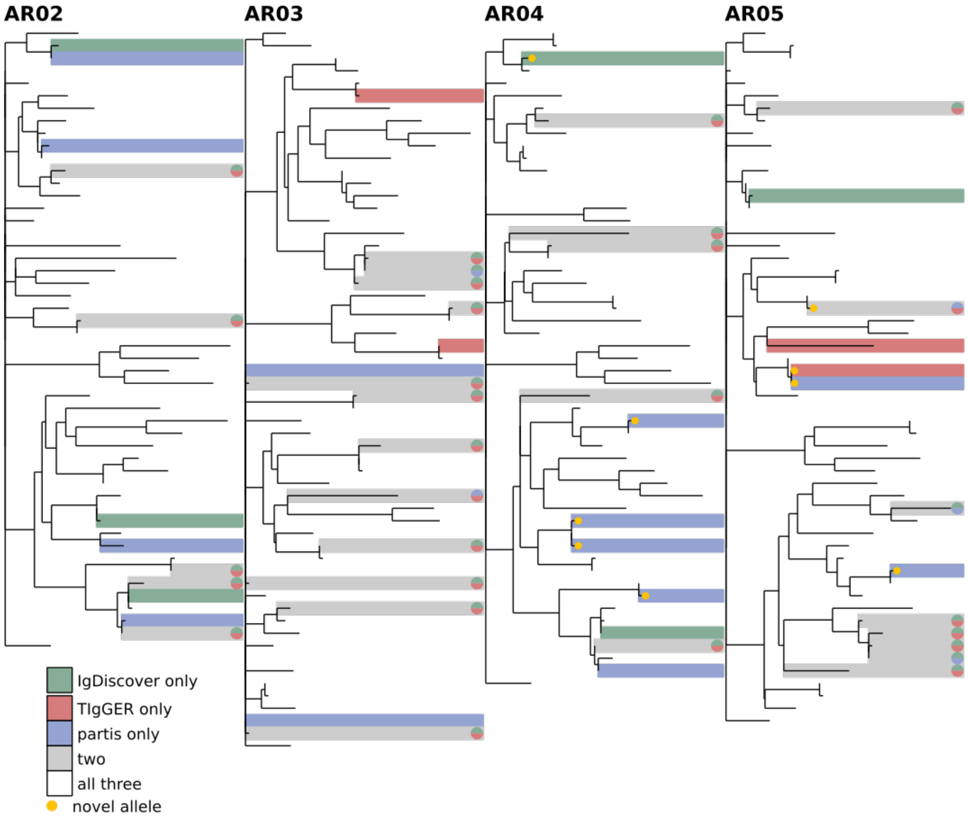

FIGURE S2. **Comparison of all three inference methods on the AR-type Myasthenia Gravis samples from [34]** (explanation in Figure 8).

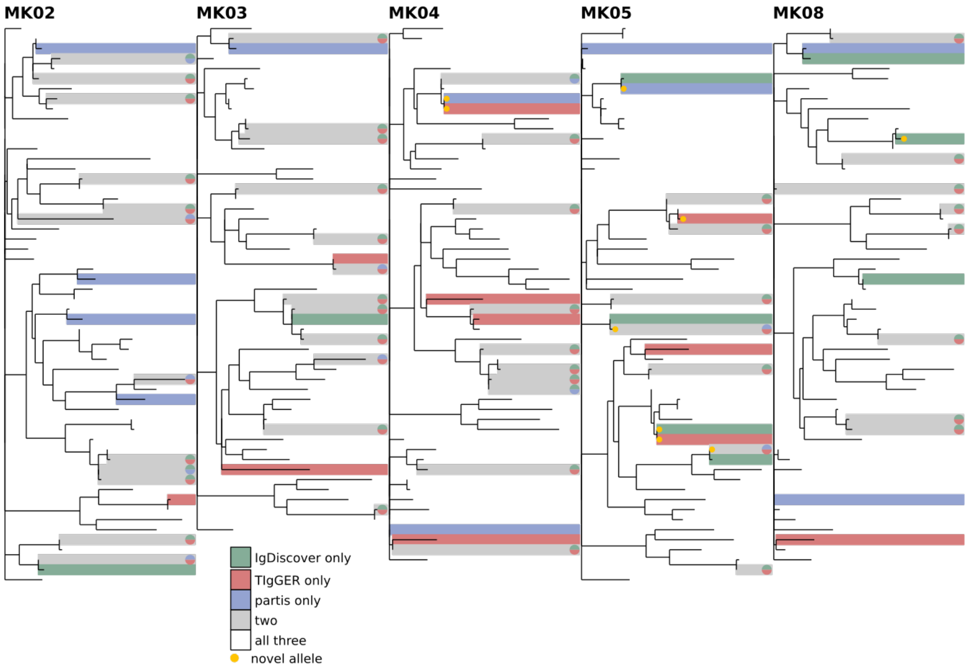

FIGURE S3. **Comparison of all three inference methods on the MK-type Myasthenia Gravis samples from [34]** (explanation in Figure 8).

Comparison between methods. While a comprehensive comparison of the details of all three methods (TIgGER, IgDiscover, and partis) is beyond the scope of this paper, we highlight some instructive details. First, the default partis method and TIgGER are more similar to each other than either is to IgDiscover. Both use a fitting procedure on mutation accumulation plots to look for new alleles. However, as described above, our approach to extracting information from these plots is very different, using hypothesis comparison rather than sharp cutoffs, for instance on the y-intercept.

IgDiscover, on the other hand, takes a quite different approach, clustering together sequences by distance and taking the consensus sequence of each cluster to be a germline gene. The main advantage of this approach is that it enables detection of new genes which are separated by either point mutations or insertions/deletions from existing alleles. The purely mutation accumulation plot-based approaches employed by TIgGER and default partis, in contrast, can only detect new alleles separated by point mutations. The tradeoff is that the fitting-based methods are able to use more detailed position-based information which allows them to function well in repertoires with higher SHM. As noted by the IgDiscover authors, distance-based clustering methods, in general, suffer from the fact that once SHM is high enough that clusters from distinct V genes start bleed together, the heuristic thresholds used to separate clusters create significant inaccuracies.

On human repertoires, since the IMGT set is already fairly complete, the ability to detect new alleles separated by insertion/deletion mutations is not particularly important. The germlines of most other species, however, are much less well characterized. It is thus quite common in these species to encounter novel alleles that are not simply allelic variants of well-known genes.

The obvious course of action, then, is to combine the mutation accumulation plot-based and clustering-based methods in order to allow accurate inference on non-naive repertoires of all species. We have, in fact, implemented this as a non-default option in partis (see Manual, at `https://git.io/vF12k`), which we recommend for non-human samples. However, after extensive validation of this combined method, we believe that a somewhat modified clustering approach will be required to achieve better performance on highly mutated samples from all species, and thus leave its description to a future paper.