Learning the heterogeneous hypermutation landscape of immunoglobulins from high-throughput repertoire data

Natanael Spisak, Aleksandra M. Walczak,* and Thierry Mora*
Laboratoire de physique de l'École normale supérieure, CNRS, PSL University, Sorbonne Université, and Université de Paris, 75005 Paris, France

Somatic hypermutations of immunoglobulin (Ig) genes occurring during affinity maturation drive B-cell receptors' ability to evolve strong binding to their antigenic targets. The landscape of these mutations is highly heterogeneous, with certain regions of the Ig gene being preferentially targeted. However, a rigorous quantification of this bias has been difficult because of phylogenetic correlations between sequences and the interference of selective forces. Here, we present an approach that corrects for these issues, and use it to learn a model of hypermutation preferences from a recently published large IgH repertoire dataset. The obtained model predicts mutation profiles accurately and in a reproducible way, including in the previously uncharacterized Complementarity Determining Region 3, revealing that both the sequence context of the mutation and its absolute position along the gene are important. In addition, we show that hypermutations occurring concomittantly along B-cell lineages tend to co-localize, suggesting a possible mechanism for accelerating affinity maturation.

I. INTRODUCTION

B cells are a crucial player in the adaptive immune system. Swift eradication of pathogens is enabled by the production of immunoglobulins (Ig) that bind tightly to antigens, helping in their detection, neutralization, and removal. Achieving high accuracy and breadth relies on the extraordinary diversity of the B cells repertoire. The process of V(D)J recombination results in a highly diverse population of naive cells [1–7]. In addition, B cells undergo affinity maturation, a Darwinian process [8] in which mutations are introduced to the immunoglobulin-coding gene and highest affinity mutants are selected [9]. This process is driven by a very high rate of somatic hypermutations (SHM), $\sim 10^{-3}$ per basepair per cell division [10], targeting the Ig genes. Some receptor genes can ultimately accumulate up to 30% amino acid substitutions, considerably altering the initial genotype. The broad diversity created by SHM ultimately ensures the emergence and selection of strong antigen binders. Understanding SHM and their statistics is key to designing better vaccination strategies [11, 12].

Like the VDJ recombination process, SHM are characterized by heterogeneous preferences. Mutational pathways affect the Ig genes unevenly, with 'cold' and 'hot' spots along the receptor gene, even before somatic selection introduces further biases [12]. SHM is initiated by Activation-Induced cytidine Deaminase (AID) through the deamination of deoxycytidines triggering an array of error-prone repair pathways [13]. AID and repair enzymes preferentially target certain regions of the gene. However, a quantitative picture of how these processes and their context dependencies result in the observed heterogeneous mutational landscape is lacking. High-throughput repertoire sequencing of the Ig gene [2, 3, 14, 15] has facilitated the development of effective models from a detailed analysis of mutational profiles of Ig sequences before [5, 16, 17] or after selection [18–22]. However, the spatial organization of mutations, their context preferences, and their interplay with selection during affinity maturation are still poorly understood, in part due to a number of confounding factors.

A fundamental issue is the bias of selection, which favors beneficial mutations over deleterious ones in the observed repertoire. This bias can be partially circumvented by analyzing synonymous substitutions [16], with the limitation that extrapolation is required to generalize to non-synonymous ones. Another way around selection is to study passenger nonproductive sequences, which are unsuccessful products of VDJ recombination and thus unaffected by selection [5, 17, 22]. These sequences make up a minority of DNA sequences, and are rarely found in mRNA sequences because of allelic exclusion, which limits their use to very large datasets.

Another confounding factor arises from phylogenetic biases due to the complex multi-lineage structure of the repertoire. While methods have been developed to infer substitution rates from lineages in a lineage-specific [21] or repertoire-wide way [23], they do not aim to correct for selection and do not address the question of hypermutation targeting.

Here we propose a new framework for quantifying and predicting immunoglobulin mutability. The model is trained on the reconstructed phylogenies of nonproductive lineages from very large published B cell repertoires totalling around half a million nonproductive sequences [7], allowing us to overcome previous limitations of dataset sizes. The approach accounts for both phylogenetic and selection biases, and allows us to study in detail the spatial and context preferences of hypermutation targeting, and to reveal the co-localization of contemporary mutations.

*Corresponding authors. These authors contributed equally.

II. RESULTS

Repertoire-wide framework to model intrinsic mutabilities from out-of-frame lineages

Out-of-frame Ig sequences are byproducts of the VDJ recombination process that are made non functional by a frameshift in the CDR3 region. Since each cell has two copies of the Ig genes, out-of-frame rearrangements may survive in the cell if recombination on the second chromosome is successful. The mechanism of allelic exclusion ensures that only the functional variant is expressed. Yet, out-of-frame IgH sequences comprise $\sim 2\%$ of rearrangements in Ig mRNA sequencing experiments, and $\sim 9\%$ in genomic DNA [6]. When a B-cell clone harboring both an out-of-frame and a functional rearrangement undergoes affinity maturation, the out-of-frame sequence acts as a passenger and mutates alongside the functional sequence, with the selection pressure acting only on the latter. While the two sequences share the same phylogeny, mutations found in out-of-frame lineages are not expected to be subject to selection.

To model the process of SHM, we reconstruct the evolutionary history giving rise to the observed mutation patterns in nonproductive rearrangements. We analysed data consisting of the IgG repertoires of 9 individuals from Ref. [7], obtained by the targeted mRNA sequencing of the Ig heavy (IgH) chain locus. We pre-processed and aligned raw IgH sequences to keep only out-of-frame sequences. We then grouped sequences into clonal families that originate from the same ancestor using single linkage clustering (Fig. 1A). The size of clonal families typically follows a power-law distribution (Fig. 1C). As a result, many lineages are represented by one or a few sequences. We focused on sufficiently large lineages (comprised of at least 6 distinct sequences) and reconstructed their lineage structure, using maximum likelihood [24, 25] to infer the topology of the underlying tree, and marginal reconstruction for the identity of ancestral states. This provides us with a list $> 200,000$ mutation events occurring between the most recent common ancestor of lineages and their leaves.

Using lineage information is essential for multiple reasons. First, it allows for a better estimate of the sequence context in which a mutation appears. In this paper we define the context as the 5-mer sequence comprising the mutated basepair flanked by 2 basepairs on each side. In the absence of lineage information, the best guess for the 5-mer context would be given by the genomic sequence of the V, D, or J segment where the mutation arose. But that context may itself be affected by other prior mutations. The tree structure allows us to identify the order of mutations and reconstruct the probable 5-mer context in which each mutation occurred. Second, for the same reason, the tree structure can help identify mutations in the hyper-variable CDR3 region, including in the junctions made of nontemplated insertions. This makes it possible to estimate the hypermutation rate in these regions. Together, these improvements mean that mutations can be identified within a broader range of 5-mer contexts, and their corresponding mutabilities better estimated. Third, lineage structure helps reduce contamination from sequences that have been under some selection. In some rare events, during affinity maturation a somatic insertion or deletion may be introduced in the CDR3 of a productive sequence, which would lead us to classify it as out-of-frame, even though it has been subject to selection prior to the frame-shift event. Focusing on mutations happening downstream of the most recent common ancestor, which is already out of frame, help us discards those contaminating events.

Given a model $P(s \to s'|t, \theta)$ of sequence evolution from s to s', where t is fraction of mutated positions between s and s', (called branch length, equal to the number of mutations divided by alignment length), and θ denotes model parameters, we can write the joint likelihood of mutation events in each lineage as

$$P(S|T, \theta) = \prod_{(i,j)\in T} P(s_i \to s_j|t_j, \theta), \tag{1}$$

where S is the set of sequences (observed and reconstructed) at each node of the tree, and T encodes the reconstructed phylogenetic tree through its branches (i, j).

We assume every position x of the sequence s evolves independently inside each branch. Mutations occur according to a set of Poisson clocks with sequence- and position-dependent rates, $\mu_{s,x}$, expressed per unit time of branch length. During t some positions will mutate and others will remain unchanged, so that

$$P(s \to s'|t) = \prod_{x \,|\, s(x)\neq s'(x)} (1 - e^{-\mu_{s,x}t}) \prod_{x \,|\, s(x)=s'(x)} e^{-\mu_{s,x}t}. \tag{2}$$

We assume that mutability depends independently on the local 5-mer sequence context centered around the mutation, $w(s, x) = (s(x-2), \ldots, s(x+2))$, and on the absolute position x along the gene (measured as the distance from the $5'$ end of the gene), so that $\mu_{s,x} = \gamma_{w(s,x)}\beta_x$. In absence of context and position dependence, we would have $\mu = 1$ by construction. Thus values of γ_w or β_x above 1 imply higher mutabilities than average, and vice versa for values below 1. To lift the degeneracy in overall scale between β_x and γ_w, we impose $\langle\beta_x\rangle = 1$.

Overall, the model has $4^5 = 1024$ parameters for γ_w corresponding to each 5-mer, and $L = 400$ parameters for β_x corresponding to each possible position. We infer these parameters from repertoire-wide sequencing data by maximizing the total log-likelihood of mutations in all branches in all lineages, $\mathcal{L}(\beta, \gamma) = \sum_{(S,T)} \ln P(S|T, \beta, \gamma)$, with respect to (β, γ), using an iterative procedure.

Validation on synthetic data

We first tested the ability of the inference framework to recover true mutability parameters using synthetic

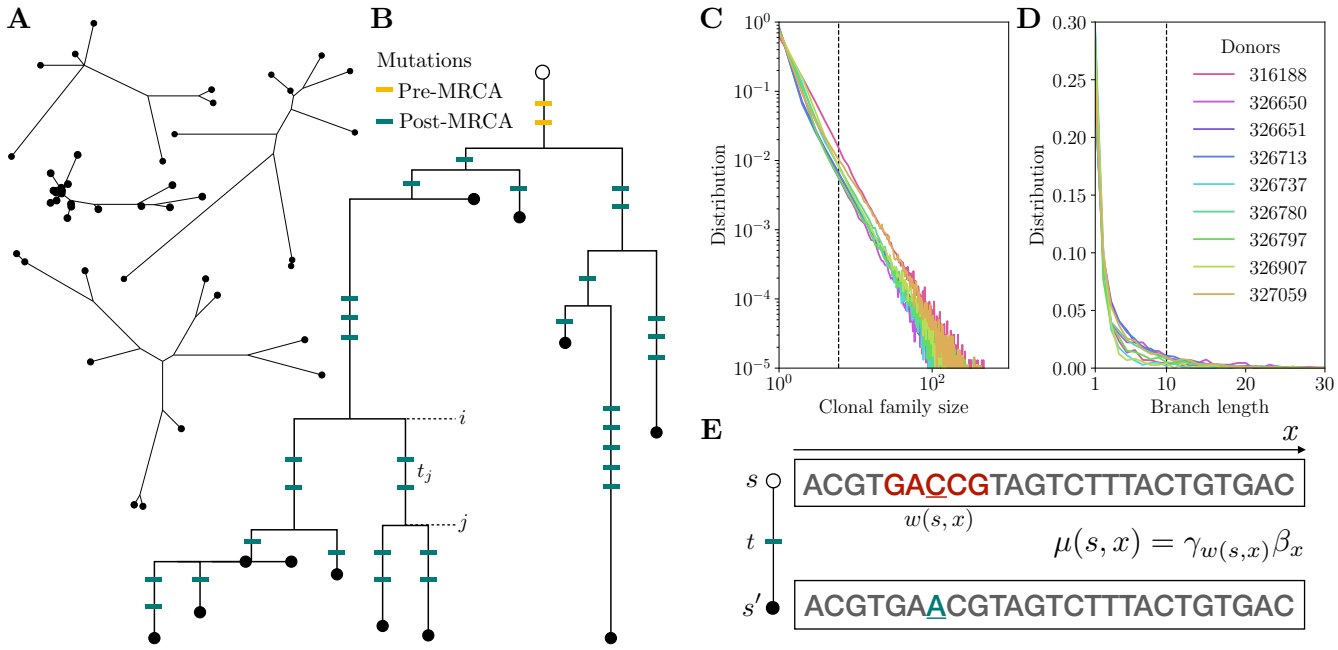

FIG. 1: Repertoire-wide framework to model somatic hypermutations (SHM) in out-of-frame lineages. A. Examples of out-of-frame clonal families. B. An example of a phylogeny with marked pre- and post-Most Common Recent Ancestor (MRCA) mutations. Only post-MRCA mutations are used for learning the somatic hypermutation model. C. Clonal family size distribution. Phylogenies were inferred for families with more than 5 unique sequences. D. Branch length distribution. Mutations encoded in branches shorter than 10 mutations were used for model inference. E. Context and position dependence of the mutation rate across a sequence. In this example the underlined C mutates to A, in the GACCG context denoted as $w(s,x)$. The mutation rate $\mu(s,x)$ depends both on the sequence s and the position x, through the context dependent rate $\gamma_{w(s,x)}$ and the position dependent rate β_x.

datasets. Synthetic data was designed to mimic as closely as possible the features of the real repertoire data to be analyzed. We used tree topologies inferred on out-of-frame lineages from 9 individuals of Ref. [7]. The sequence at the root of each tree was replaced by a random sequence drawn using IGoR, a model of stochastic VDJ recombination [22]. Random mutations were then introduced along the tree structure, following the same number of mutations on each branch as in the original lineage, and according to the SHM model (Eq. 2). Context-dependent parameters γ_w were set to the previously published S5F model [16] , and three variants of the position dependent β_x were tested: flat, and two sinusoidal profiles (see Methods). Finally we collected sequences at the leaves of the trees into a synthetic dataset.

Starting from this dataset, we performed alignment, clonal family inference, tree reconstruction and finally model inference using the exact same procedure as for real data. We compared parameters inferred this way to the true values of γ_w and β_x (Fig. 2). We were able to recover these rates with excellent accuracy (Pearson's $r^2 = 97\%$ for both γ and β).

The fact that the procedure recovers the correct position-dependent profile β_x, including a flat one (Fig. 2C), shows that the framework successfully corrects for the two following confounding factors. First, sequence conservation across the different V, D, and J segments means that context and position are often intertwined, making the extraction of each dependence difficult. Second, high variability in the CDR3 may cause errors in the assignment of sequences into clonal families, and makes it harder to reliably call mutations than in the germline regions. This remains true in the presence of large variations of the mutability along the position, including in the CDR3, as demonstrated on the sinusoidal profiles (Fig. 2D). On the other hand, the possibility to use the CDR3 sequence for model inference gives access to a more diverse range of possible contexts, leading to better estimates for contexts that are underrepresented in the germline genes.

To assess the impact of errors in the reconstruction of clonal families and lineages on the inferred parameters, we repeated the procedure using the true tree topologies instead of the reconstructed ones. This only modestly improved accuracy ($r^2 = 98\%$, see Fig. S1), suggesting that the procedure is robust to lineage misassignments.

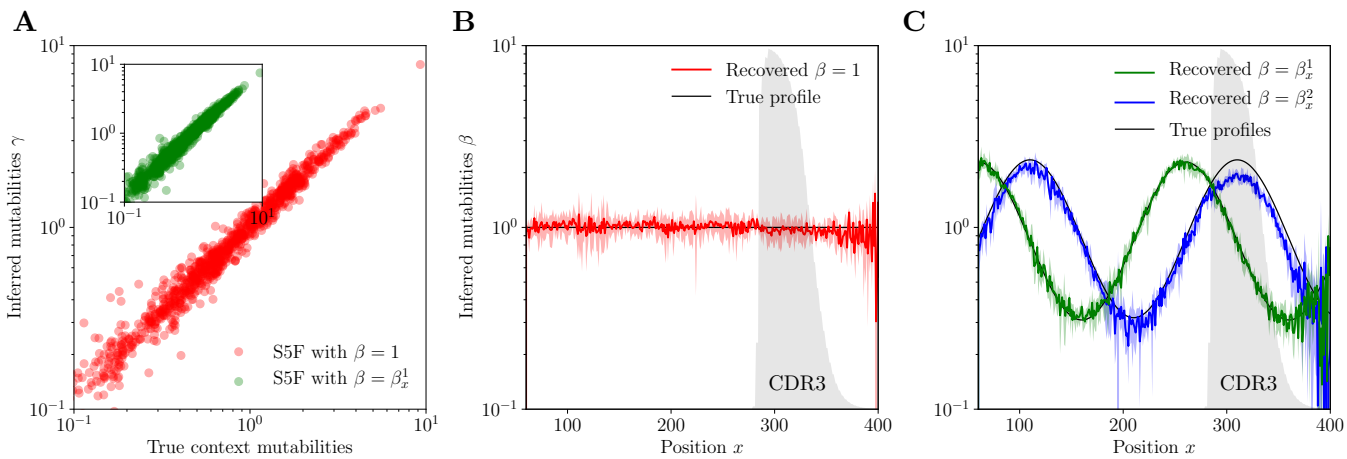

FIG. 2: Validation of the SHM model inference framework on synthetic data generated with the S5F model [16]. A. Inference of context mutabilities γ_w. B,C: Inference of position mutabilities β_x for flat (B) and sinusoidal (C) profiles. Error bars correspond to 95% confidence intervals. The frequency at which a given position belongs to the CDR3 region is indicated with the grey shaded area.

Mutabilities depend on both sequence context and position

Confident that our procedure is able to infer rates reliably, we next applied it to real data, consisting of the out-of-frame lineages from Ref. [7]. The inferred dependencies of mutability with context and position are presented in Fig. 3. We represent context dependence using a flat variant of the "hedgehog" plots used in Ref. [16], for A-, T- , C-, and G-centered motifs (Fig. 3A-D). Full parameter tables are available at https://github.com/statbiophys/shmoof.

Context dependent rates for A-centered motifs correspond well to the standard W$\underline{\text{A}}$ classification [26]: 76% of A-centered 5-mers with $\gamma_w > 1$ are of the W$\underline{\text{A}}$ type, and only 7 of 128 W$\underline{\text{A}}$ 5-mers have $\gamma_w \lesssim 1$. T-centered motifs are dominated by coldspots and their mutabilities do not align well with their corresponding reverse complement counterparts. This is in agreement with the known property of Polymerase η to be prone to errors at A nucleotides on the top strand [27].

The C- and G-centered motifs have largely reverse-complement-symmetric rates (see Fig. S2). As previously noted [16], this is in agreement with the strand-symmetric targeting of C/G-centered motifs by the AID enzyme.

The previously reported WR$\underline{\text{C}}$Y/R$\underline{\text{G}}$YW motif [13, 28] predicts high mutability reasonably well, while the SY$\underline{\text{C}}$/$\underline{\text{G}}$RW class of motifs [29] explains well a good fraction of coldspot motifs. Importantly, a large number of high or low mutability 5-mers do not belong to any of the previously reported motifs (see Supplementary Tables 1 and 2).

The rugged profile of position dependence (Fig. 3E) shows clear enrichment in mutations in the CDR1 and CDR2 regions, reflected in the up to 2-fold increase of the position-dependent rates. Framework regions are less mutated and we also observe a slight drop in the mutabilities of the positions beyond the Cysteine anchor of the CDR3 region. We also learned models where the position was defined from the 3' end of the sequence in the J segment (Fig. S3), yielding similar results but no clear improvement over 5'-based position. High mutability of CDR1 and CDR2 was already noted [30] and justified as an enrichment in highly mutable motifs (as quantified with the S5F model). Our findings suggest that there is a secondary mechanism of this enrichment, having to do either with accessibility of mutation-inducing enzymes or a superposition of context-dependent effects that evade the assumption of independent evolution at different sites and the limitation of 5-mer motifs.

Note that introducing the explicit position dependence does affect the learning of the context-dependent parameters: learning γ_w with no position dependence (fixing $\beta_x = 1$) yields similar but markedly different parameters than when learning a free β_x ($r^2 = 81\%$, Fig. S4).

Model is consistent across individuals and explains data better than previous approaches

To check the model's generality, we estimated its variability across individuals by computing Pearson's correlation coefficient between the context (γ_w, Fig. 4A) and position (β_x, Fig. 4B) mutability profiles of different donors. The precision with which we can estimate model parameters depends on the number of sequences used for inference, particularly for rare 5-mer contexts. Because two individuals had many more reads than the others 7, we pooled together these seven individuals to make com-

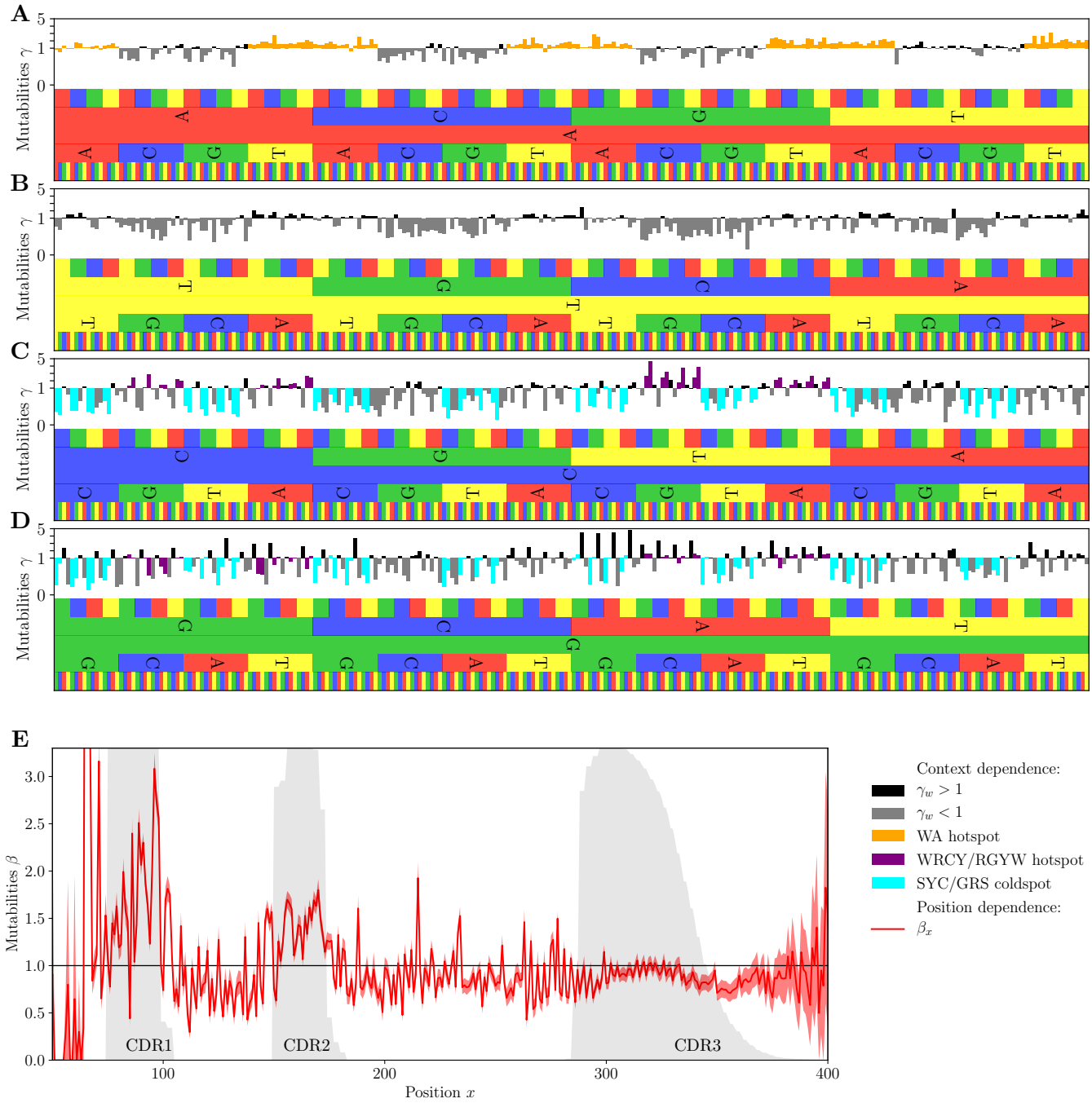

FIG. 3: Context and position dependent model parameters. Context-dependent mutabilities γ_w for A- (A),T- (B),C- (C) and G-centered (D) 5-mers. The colors indicate known hopspot and coldspot motifs. E. Position-dependent mutabilities β_x. Gray shadings show the probability to be in the CDR1, CDR2, or CDR3 regions. Error bars correspond to 95% confidence intervals. See Fig. S5 for an full analysis of parameter uncertainty.

parisons with similar dataset sizes (Fig. 4C). We then compared the 2 individuals and 1 meta-individual with each other and with a model learned on data from all individuals. For the 2 individuals with the largest repertoire datasets, the results are highly reproducible with Pearson's $r^2 = 78\%$ for context and $r^2 = 70\%$ for position parameters (Fig. 4A), suggesting that the model captures universal biochemical properties of the hypermutation process.

To further validate the model's accuracy, we compared its prediction to data on the V-specific mutation profiles, which consist of the position-dependent mutation rate

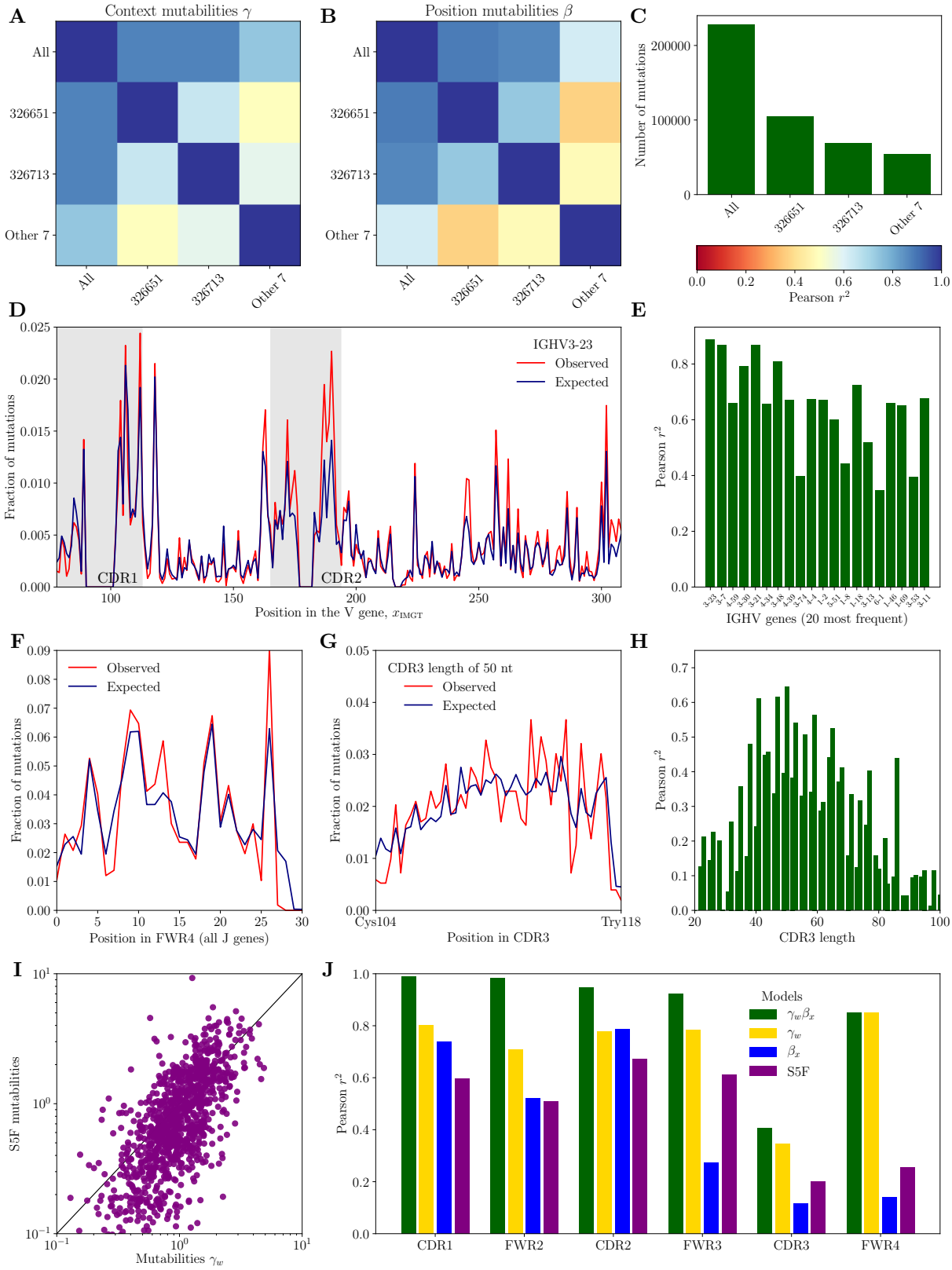

FIG. 4: The model explains the data. Observed profiles were measured across the entire dataset used for model inference. See Fig. S6 for an equivalent figure when data was divided into training and testing sets. Reproducibility of parameters for individual-specific models: context (A) and position (B) mutabilities. C. Number of mutations used for inference D. An example mutation profile in the most common V gene. E. Model performance across V genes. F. Mutation profile in the FWR4 region. G. An example mutation profile in the CDR3 region for CDR3 length of 50 nts. H. Model performance across CDR3 lengths. I. Comparison with the S5F model. J. Summary of models performance across sequence regions. We compare the full context- and position-dependent model ($\gamma_w\beta_x$) with purely context- (γ_w) and position- (β_x) dependent models as well as with the S5F default model.

for each V segment. These rates result from the combined effect of position and context, but they are not fitted directly by the model. A typically good example of such a profile is shown in Fig. 4D. The prediction is generally excellent (Pearson's $r^2 \sim 50 - 80\%$), and is poorest for V segments for which little data was available (Fig. 4E). Similarly, the model predicts well the mutability on Framework Region 4 (FWR4), which encompasses the J segment (Fig. 4F), as well as in the CDR3 (Fig. 4G and H), which is usually ignored in other approaches. Performance is best for the most frequent CDR3 length (Fig. 4H).

We compared the results of our inference to the S5F model [16], which was trained on independent data. The S5F model is defined by a mutability table γ_w with no attempt to disentangle position dependence, so a direct comparison is subject to caution. Besides, S5F mutabilities are learned from synonymous mutations of productive sequences, requiring extrapolation methods to cover all 1024 contexts, all of which do not occur with synonymous mutations. Yet, the two sets of mutabilities γ_w correlate fairly well ($r^2 = 36\%$, Fig. 4I). Correlation rises to $r^2 = 44\%$ for contexts appearing in synonymous mutations, versus $r^2 = 18\%$ for the other contexts for which S5F recourses to extrapolation, emphasizing the limitations of that extrapolation.

A summary of the performances of the different modeling approaches on the mutabilities in the different regions of the IgH gene is shown in Fig. 4J. We also checked for overfitting by dividing the dataset into a training and a testing, finding similar results (Fig. S6). The full position and context dependent model ($\mu_{s,x} = \gamma_w \beta_x$) performs better than models with context or position alone ($\mu_{s,x} = \gamma_w$ and $\mu_{s,x} = \beta_x$). While the context explains the bulk of the mutation profile, adding positional effects substantially boosts performance. Our model clearly outperforms the S5F model, although it should be reminded that S5F was trained on a distinct dataset. Re-training S5F on the productive sequences from the present datasets using the procedure described in the original article [16] actually yielded worse performance (data not shown), for reasons that are unclear to us. Overall, accounting for phylogeny and disentangling the combined effects of context and position allows our model to accurately predict mutabilities including in the hyper-variable CDR3 region.

Co-localization of mutations cannot be explained by context and position bias

It was previously observed that hypermutations tend to cluster along genomic position in nonproductive sequences [22]. However, the origin of this phenomenon and its dependence on confounding factors such as phylogeny and heterogeneous hot spot concentration were not fully characterized.

Clustering of mutations can be directly observed by plotting the fraction $n(r)$ of pairs of mutations at distance r from each other as a function r (normalized by the total number of pairs at that distance, see Methods), which is also called a spatial correlation function in physics. Focusing on lineages with at least 6 leaves, and iterating through all branches with fewer than 10 mutations, we evaluated this correlation function for pairs of mutations occurring in the same branch of the phylogeny *versus* distant branches, as schematized in Fig. 5A). We then compared this correlation function to our model predictions (Fig. S7). The enrichment of closeby mutations can be quantified by the normalized correlation function $f(r) = n(r)/n_m(r)$, where $n(r)$ stands for the fraction of pairs of mutations distant by r in the same tree branch and $n_m(r)$ the model prediction (Fig. 5B).

Pairs of mutations in distinct branches are well explained by the model, suggesting that they are independent of each other, in agreement with the biological picture that they occur at different rounds of affinity maturation. The enrichment of closeby mutations in distant branches can be entirely explained by the clustering of hotspot regions. Interestingly, both context and position dependencies of the mutability are needed to explain the data (Fig. S8). In contrast, pairs of mutations inside branches tend to occur closer to each other than predicted by the model. The enrichment of closeby mutations is up to five-fold, pointing to an additional mechanism causing hypermutation clustering. We observe that this enrichment persists in the presence of selection, as verified by computing the correlation function $f(r)$ in productive lineages (Fig. 5C).

Minimal model of co-localization

To explain the observed excess of co-localized mutations, we propose a simple phenomenological model. Targeted mutations, following the context and position dependent profiles described so far, cause additional nearby 'follow-up' mutations due to error-prone DNA repair. Given a substitution at x_0 drawn from the same distribution as before, each position $x \neq x_0$ can subsequently mutate with probability

$$p(x|x_0) = \epsilon\, e^{-|x-x_0|/\xi}, \tag{3}$$

where ξ is the correlation length and ϵ is small. The total number of follow-up mutations is approximately Poisson distributed with mean $\sum_x p(x|x_0) \simeq 2\epsilon/(1 - e^{-1/\xi})$. To simulate this process, we followed the same procedure as described earlier for synthetic data, but allowing for follow-up, as well as targeted mutations, while keeping the total number of mutations in each branch constant. We then computed the correlation function $f(r)$, and compared it to true profiles (Fig. 5B). We obtain a good agreement for $\xi = 10$ and $\epsilon = 5\%$ corresponding to an average of ~ 1 follow-up mutation per targeting event. This result suggests that as many as 50% of observed mutations are follow-up mutations.

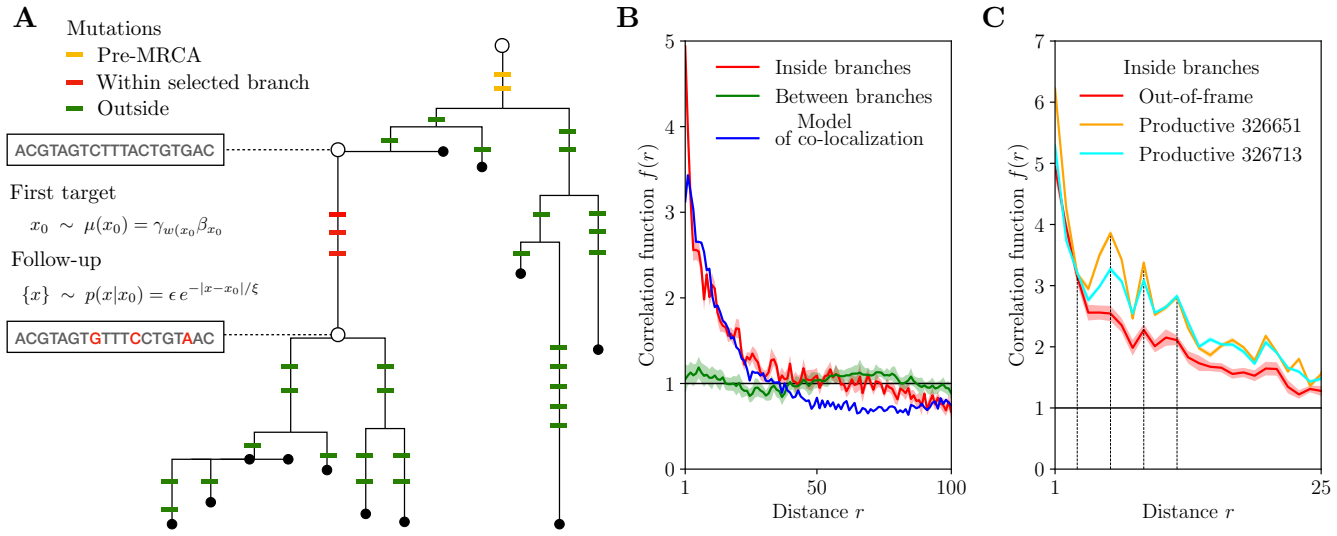

FIG. 5: Co-localization of subsequent hypermutations. A. Co-localization model explained on an example phylogeny. An initial mutation is drawn from the context- and position-dependent model. Then, follow-up mutations are drawn in its vicinity within the same branch. B. Correlation function $f(r)$ for pairs of mutations: inside (red) and between (green) branches and for simulated mutations according to the co-localization model (3) with $\epsilon = 5\%$ and $\xi = 10$. C. Co-localization of subsequent hypermutations in productive lineages from the 2 largest individuals' datasets (ID=326651 and ID=326713). Correlation function $f(r)$ for pairs of mutations inside branches compared to the out-of-frame result. The first multiplicities of the codon frame length, $r = 3, 6, 9, 12$ are marked with dotted lines to guide the eye. Shaded areas represent 95% confidence intervals.

III. DISCUSSION

The mutational landscape of antibody repertoires results from many entangled effects, which are often lumped together into effective models of hypermutations [12, 16, 31]. First, hypermutations have intrinsic preferences for certain positions along the IgH gene, regardless of their impact on protein function. In addition, selection for antibody function, which includes protein stability and antigen affinity, favors beneficial mutations and suppresses deleterious ones [13]. While intrinsic SHM preferences are believed to be universal, selective forces vary across lineages which are involved in distinct immune responses [21], and may also depend on the individual's immune status [32]. Repertoire sequencing gives a snapshot of a rapidly adapting population subjected to these forces, making it hard to disentangle intrinsic SHM preferences from the combined effects of selection and genetic drift. By focusing on non-productive lineages and using a phylogeny-based approach, we overcome the biases arising from the dynamics of affinity maturation to obtain a comprehensive picture of SHM intrinsic preferences.

Each hypermutation occurs through a series of events of DNA damage and repair. The action of each enzyme, including AID to error-prone DNA repair enzymes, may each have their own sequence preferences, and the interplay of these different biases results in the observed profile. In our approach, these complex mutational pathways are subsumed into an effective model with a limited number of interpretable parameters in terms of effective context and position dependence. As a result, the context dependent weights γ_w do not simply reflect the binding preference of AID, but also account for the biases of other biochemical steps. Our framework enables direct measurement of the mutability γ_w of a wide range of 5-mer contexts, recovering the known classifications of hot and cold spots [16, 28]. We show that our model outperforms existing methods as well as purely context or position dependent models in terms of explaining the data.

The introduction of an explicit and universal position dependence, β_x, allows us to unveil an excess of mutations in the CDR1 and CDR2 regions. This enrichment of mutations cannot be simply explained by their harboring more hotspot contexts. We cannot exclude that this residual position dependence is due to more complex context effects missed by our model (based e.g. on 7-mers, which would be impractical to infer from the present dataset). Alternatively, SHM may preferentially target these regions independently of their sequence context, possibly through epigenetic mechanisms. Such preference is known to exist at the genome-wide level to mutate the Ig loci without affecting other genes [13], so it is plausible that the same mechanism targets some specific positions within Ig. The enrichment of mutations in the CDR1 and CDR2 regions is even more marked in productive sequences, meaning that these mutations are more likely to be selected during affinity maturation. This suggests that the intrinsically enhanced mutability of these regions may carry an evolutionary advantage, by focusing hypermutations on regions where they are the

most beneficial [30]. The stability of the immunoglobulin relies on the FWR regions and most of the substitutions are expected to be deleterious. The purifying nature of selection in FWR regions has been quantified in Ref. [33] and contrasted with positive selection in CDR regions.

By studying mutations along lineages, we were able to study mutations in the probable context in which they occurred, rather than relative to the germline sequence, allowing us to take into account the order of mutations and to sample a broader diversity of 5-mer contexts. This approach also allowed us to study and characterize hypermutations in the CDR3, which has been neglected in previous work [11] owing to the difficulty to separate these mutations from junctional diversity.

The phylogenetic methods employed in this study were not specifically designed to study B cell repertoires. In particular the assumptions allowing for fast likelihood computations do not account for the context dependence of the mutation rate beyond the codon frame [23]. The position-dependent model introduced here could offer a compromise. While it does not account for the the full complexity of SHM biases, it captures the variation of the mutation rate observed in out-of-frame data well (Fig. 4), and can operate under the assumption of independent site evolution. Our framework could also be easily extended to include position-dependent selection in the nucleotide or amino acid representation.

Our analysis confirmed a phenomenon of co-localization of mutations along the sequence. While this effect had been previously reported [22], here we showed that it could not be explained by phylogenetic bias or the existence of regions of higher and lower mutabilities. We proposed a minimal quantitative model of hypermutation targeting, followed by error-prone DNA repair that causes follow-up mutations, which explains the data well. While ideally we would like to infer the position and context mutability profiles taking these follow-up mutations into account, the task is impractical because it would require to identify the origin of each mutation. We expect that doing so would only renormalize the values of the context preferences. While the adaptive advantage of co-localized mutations is unclear, we find the correlation function in productive lineages follows the unproductive baseline with additional enrichment enhanced at multiples of the codon length, 3, suggesting signatures of selection (Fig. 5C). We speculate that nearby mutations occurring simultaneously could help cross barriers of positive sign epistasis, whereby two or more mutations are deleterious by themselves, but beneficial together. This phenomenon could accelerate affinity maturation by favoring compensatory or epistatic mutations at amino acids that interact strongly within the antibody protein [34, 35].

The obtained mutability models make predictions about the likelihood and plausibility of particular trajectories of affinity maturation. They could be useful in designing vaccination strategy, by helping choose targets with a greater potential for accumulating beneficial mutations towards antibodies with desired properties such as neutralization power, or broadness in the case of fast evolving pathogens such as influenza or HIV [11, 36].

IV. METHODS

Data and preprocessing

We perform the analysis on recently published high-throughput RNA sequencing of Ig heavy genes at great depth [7].

The sequences were barcoded with unique molecular identifiers (UMI) to correct for the PCR amplification bias. However, UMI cannot be used to correct sequencing errors, as most UMI were represented by a single sequence: the number of UMIs used is of the same order as the total number of cells in use. We aligned raw sequences using presto of the Immcantation pipeline [37] with setup allowing to correct for errors in UMIs and deal with insufficient UMI diversity. The V region primers were masked and the C region primers were used to distinguish the two isotypes of sampled B cells: the IgM and IgG classes. The study of mutation profiles in the two groups revealed a much lower mutational load in the IgM cohort and hence a higher relative level of sequencing errors, as well as shallower tree topologies. For further analysis we chose to focus exclusively on the IgG class. Reads were filtered for quality and paired using default presto parameters. Pre-processed data was then aligned to V, D and J templates from IMGT [38] database using IgBlast [39]. In total there were 3.6×10^6 IgG sequences per person (average 3.6×10^6, median 1.8×10^6), of which up to 2% were unproductive (average 5.7×10^4, median 2.9×10^4).

Inference of evolutionary trajectories

Sequences with a frameshift in the CDR3 region were then selected and used to reconstruct clonal families as follows. In the first step, reads were aligned to the V and J templates and grouped into classes of sequences with the same V and J gene assigned, as well as equal CDR3 length. In the out-of-frame classes we inferred clonal lineages by single linkage clustering with a threshold of 90% on CDR3 region identity [40]. We reconstruct maximum likelihood topologies, as well as the identity of ancestral states, under a simple K80 model of character evolution [41] for all lineages comprising at least 6 unique sequences. The model does not capture the complexity of the observed mutation profile, but avoids fitting multiple parameters independently in small lineages of relatively short alignment. The existing repertoire-wide method [23] is incompatible with out-of-frame lineages since it operates on 61 productive codons. Ancestral states are found through marginal reconstruction. Germline V and

J sequences were used as an outgroup to inform the phylogenetic inference and root the lineage.

Model inference

With the exception of the initial branch, which joins the germline sequence and the most recent common ancestor of the lineage, all branches shorter than 10 substitutions were used for model inference.

Our task is to find a set of parameters $\{\gamma_w\}, \{\beta_x\}$ that maximise the log-likelihood

$$\mathcal{L} = \sum_{S,T;(i,j)\in T} \left[\sum_{x' \mid s_i(x') \neq s_j(x')} \ln\left(e^{\gamma_{w(s_i,x')}\beta_{x'}t} - 1\right) - \sum_x \gamma_{w(s_i,x)}\beta_x t \right], \tag{4}$$

where S is the set of sequences (observed and reconstructed) at each node of the tree, and T encodes the reconstructed phylogenetic tree through its branches (i,j), with reconstructed ancestral states s_i and s_j. The rates $\mu_{s,x} = \beta_x \gamma_{w(s,x)}$ are defined so that the length of each branch t is expressed in terms of the expected number of substitutions per basepair (total number of substitutions divided by the total alignment length). Imposing $\partial\mathcal{L}/\partial\gamma_w = 0$ yields an implicit expression for γ_w as a function of $\{\beta_x\}$, but independent of $\{\gamma_{w'}\}_{w'\neq w}$, which can be solved by one-dimensional root finding. Likewise, setting $\partial\mathcal{L}/\partial\beta_x = 0$ gives an implicit expression for β_x as a function of $\{\gamma_w\}$. We can perform the following iteration:

$$\gamma^n = \arg\max_\gamma \mathcal{L}(\gamma, \beta^n) \tag{5}$$

$$\beta^{n+1} = \arg\max_\beta \mathcal{L}(\gamma^n, \beta), \tag{6}$$

which converges to the maximum of $\mathcal{L}$ with respect to the joint $\{\gamma_w, \beta_x\}$.

To estimate the uncertainty of inferred parameters we sample with replacement from the set of all branches to create 400 bootstrap copies. We report 95% confidence intervals.

Substitution models

Not only the targeting rate, but also the identity of the substitution is known to depend on the identity of neighboring bases [16]. In our formulation of the model, inference of the targeting rates does not require knowing the substitution type, however we can easily extend the framework to include this dependence. The probability of mutating from w to w' over a period t can be expressed as

$$P(w \to w'|t) = 1 - e^{-\gamma_w \omega_{w\to w'} \beta_x t}, \tag{7}$$

where $\sum_{w'} \omega_{w\to w'} = 1$ and $\omega_{w\to w'} \neq 0$ if w' is a result of a substitution at the central position of w. This way we add $2 \times 4^5 = 2048$ parameters to the model. We can infer the maximum likelihood estimates of $\omega_{w\to w'}$ using the same iterative scheme introduced in the previous section.

Synthetic datasets

We created synthetic datasets using the S5F model of mutability (downloaded from clip.med.yale.edu/shm) for γ_w. We used a flat profile, $\beta_x = 1$ as well as sinusoidal profiles $\ln \beta_x^1 = 2\sin(x/\delta) - 1$ and $\ln \beta_x^2 = 2\cos(x/\delta) - 1$ with $\delta = 50$. For each branch (i,j), we compute the mutability $\mu_{s_i,x}$ as a function of x, and then introduce mutations at n random positions picked without replacement according to $\mu_x / \sum_{x'} \mu_{x'}$, where n is the number of mutations on the branch (fixed by the lineage structure taken from the real data).

Data availability

Code for producing the figures of this paper, as well as the inferred model parameters, are freely available at https://github.com/statbiophys/shmoof.

[1] Hozumi N, Tonegawa S (1976) Evidence for somatic rearrangement of immunoglobulin genes coding for variable and constant regions. *Proceedings of the National Academy of Sciences* 73:3628–3632.

[2] Boyd SD, et al. (2009) Measurement and clinical monitoring of human lymphocyte clonality by massively parallel {VDJ} pyrosequencing. *Sci Transl Med* 1:12ra23.

[3] Glanville J, et al. (2009) Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proceedings of the National Academy of Sciences* 106:20216–20221.

[4] Larimore K, McCormick MW, Robins HS, Greenberg PD (2012) Shaping of Human Germline {IgH} Repertoires Revealed by Deep Sequencing. *The Journal of Immunology* 189:3221–3230.

[5] Elhanati Y, et al. (2015) Inferring processes underlying B-cell repertoire diversity. *Philos Trans R Soc Lond, B, Biol Sci* 370:20140243.

[6] DeWitt WS, et al. (2016) A Public Database of Memory and Naive B-Cell Receptor Sequences. *Plos One* 11:e0160853.

[7] Briney B, Inderbitzin A, Joyce C, Burton DR (2019) Commonality despite exceptional diversity in the baseline human antibody repertoire. *Nature* 566:393–397.

[8] Cobey S, Wilson P, Iv FAM, Cobey S (2015) The evolution within us.

[9] Mesin L, Ersching J, Victora GD (2016) Germinal Center B Cell Dynamics. *Immunity* 45:471–482.

[10] Kleinstein SH, Louzoun Y, Shlomchik MJ (2003) Estimating hypermutation rates from clonal tree data. *The Journal of Immunology* 171:4639–4649.

[11] Bonsignori M, et al. (2016) Maturation pathway from germline to broad hiv-1 neutralizer of a cd4-mimic antibody. *Cell* 165:449 – 463.

[12] Schramm CA, Douek DC (2018) Beyond hot spots: Biases in antibody somatic hypermutation and implications for vaccine design. *Frontiers in Immunology* 9:1–11.

[13] Feng Y, Seija N, Di Noia JM, Martin A (2020) AID in Antibody Diversification: There and Back Again. *Trends in Immunology* pp 1–15.

[14] Weinstein JA, Jiang N, White RA, Fisher DS, Quake SR (2009) High-throughput sequencing of the zebrafish antibody repertoire. *Science* 324:807–810.

[15] Reddy ST, et al. (2010) Monoclonal antibodies isolated without screening by analyzing the variable-gene repertoire of plasma cells. *Nat Biotechnol* 28:965–969.

[16] Yaari G, et al. (2013) Models of somatic hypermutation targeting and substitution based on synonymous mutations from high-throughput Immunoglobulin sequencing data. *Frontiers in Immunology* 4.

[17] Cui A, et al. (2016) A Model of Somatic Hypermutation Targeting in Mice Based on High-Throughput Ig Sequencing Data. *The Journal of Immunology* 197:3566–3574.

[18] Mccoy CO, et al. (2015) Quantifying evolutionary constraints on B-cell affinity maturation. *Philos Trans R Soc Lond, B, Biol Sci* 370:20140244.

[19] Sheng Z, et al. (2017) Gene-specific substitution profiles describe the types and frequencies of amino acid changes during antibody somatic hypermutation. *Frontiers in Immunology* 8:1–14.

[20] Hoehn KB, Lunter G, Pybus OG (2017) A Phylogenetic Codon Substitution Model for Antibody Lineages. *Genetics* 206:417–427.

[21] Dhar A, Davidsen K, Matsen FA, Minin VN (2018) Predicting B Cell Receptor Substitution Profiles Using Public Repertoire Data.

[22] Marcou Q, Mora T, Walczak AM (2018) High-throughput immune repertoire analysis with IGoR. *Nature Communications* 9:561.

[23] Hoehn KB, et al. (2019) Repertoire-wide phylogenetic models of B cell molecular evolution reveal evolutionary signatures of aging and vaccination. *Proceedings of the National Academy of Sciences of the United States of America* 116:22664–22672.

[24] Felsenstein J (1981) Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution* 17:368–376 Publisher: Springer-Verlag.

[25] Stamatakis A (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312–1313 arXiv: 10.1093/bioinformatics/btu033 ISBN: 1367-4811.

[26] Zhao Y, et al. (2013) Mechanism of somatic hypermutation at the wa motif by human dna polymerase eta. *Proceedings of the National Academy of Sciences* 110:8146–8151.

[27] Pilzecker B, Jacobs H (2019) Mutating for good: Dna damage responses during somatic hypermutation. *Frontiers in Immunology* 10:438.

[28] Unniraman S, Schatz DG (2007) Strand-biased spreading of mutations during somatic hypermutation. *Science* 317:1227–1230.

[29] Pham P, Bransteitter R, Petruska J, Goodman MF (2003) Processive aid-catalysed cytosine deamination on single-stranded dna simulates somatic hypermutation. *Nature* 424:103–107.

[30] Saini J, Hershberg U (2015) B cell variable genes have evolved their codon usage to focus the targeted patterns of somatic mutation on the complementarity determining regions. *Molecular Immunology* 65:157 – 167.

[31] Dhar A, Ralph DK, Minin VN, IV FAM (2019) A bayesian phylogenetic hidden markov model for b cell receptor sequence analysis.

[32] Zuckerman NS, et al. (2010) Somatic hypermutation and antigen-driven selection of b cells are altered in autoimmune diseases. *Journal of Autoimmunity* 35:325 – 335.

[33] Nourmohammad A, Otwinowski J, Łuksza M, Mora T, Walczak AM (2019) Fierce Selection and Interference in B-Cell Repertoire Response to Chronic HIV-1. *Molecular Biology and Evolution* 36:2184–2194.

[34] Koenig P, et al. (2015) Deep sequencing-guided design of a high affinity dual specificity antibody to target two angiogenic factors in neovascular age-related macular degeneration. *Journal of Biological Chemistry* 290:21773–21786.

[35] Adams RM, Kinney JB, Walczak AM, Mora T (2019) Epistasis in a Fitness Landscape Defined by Antibody-Antigen Binding Free Energy. *Cell Systems* pp 1–8.

[36] Liao HX, et al. (2013) Co-evolution of a broadly neutralizing HIV-1 antibody and founder virus. *Nature* 496:469–476.

[37] Vander Heiden JA, et al. (2014) pRESTO: a toolkit for processing high-throughput sequencing raw reads of lymphocyte receptor repertoires. *Bioinformatics* 30:1930–1932.

[38] Giudicelli V, et al. (2006) IMGT/LIGM-DB, the IMGT® comprehensive database of immunoglobulin and T cell receptor nucleotide sequences. *Nucleic Acids Research* 34:D781–D784.

[39] Ye J, Ma N, Madden TL, Ostell JM (2013) IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Research* 41:W34–W40.

[40] Gupta NT, et al. (2015) Change-O: A toolkit for analyzing large-scale B cell immunoglobulin repertoire sequencing data. *Bioinformatics* 31:3356–3358.

[41] Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* 16:111–120.

Supplementary information

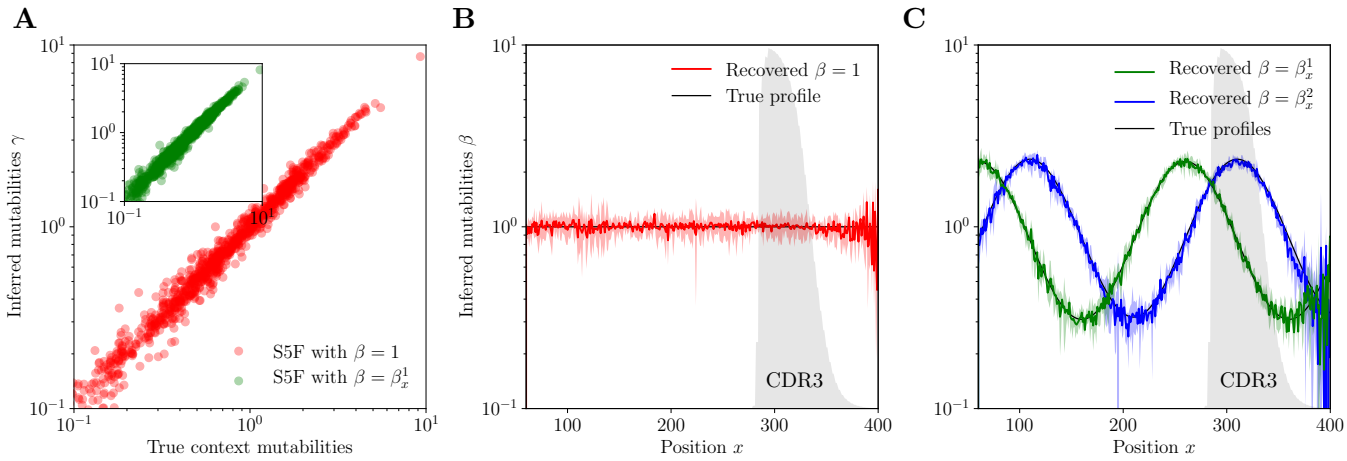

FIG. S1: Model inference on synthetic data using true phylogenies. A. Inference of context mutabilities γ. B,C. Inference of position mutabilities β for flat and sinusoidal profiles. Error bars correspond to 95% confidence intervals. Frequency at which a given position belongs to the CDR3 region is indicated with the grey shaded areas.

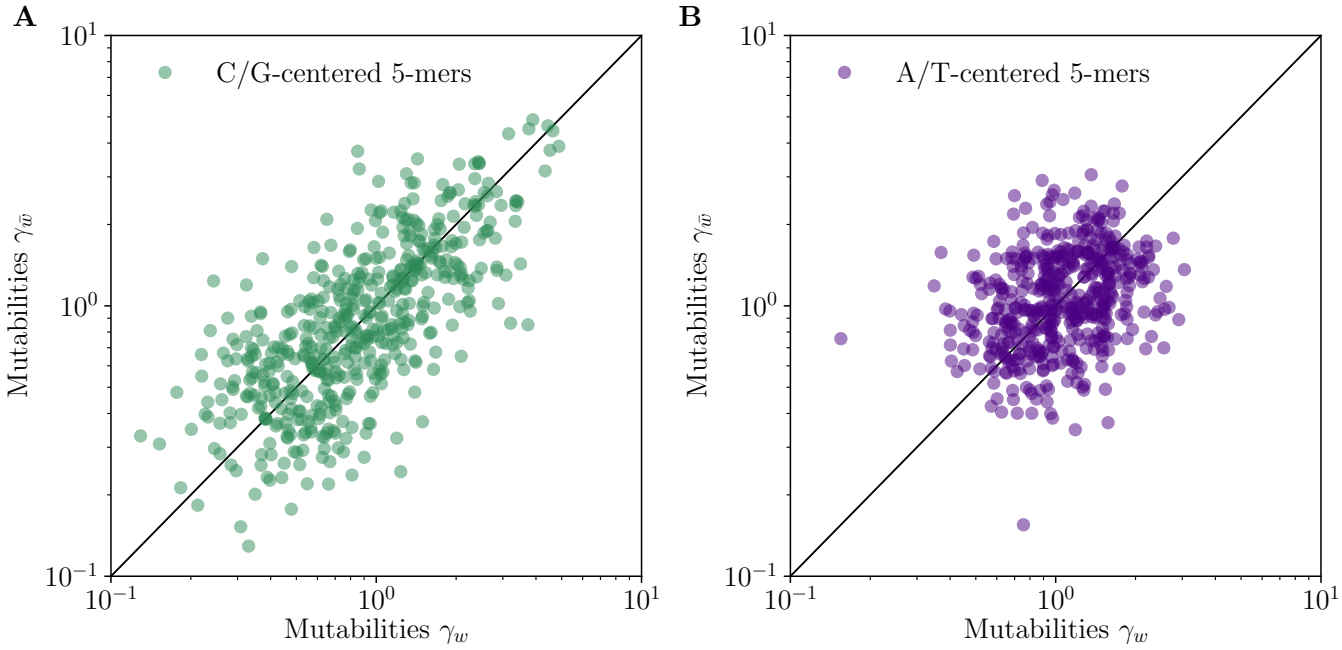

FIG. S2: Strand asymmetry of the context-dependent rates. For each motif w we juxtapose its mutability γ_w with the mutability of its reverse complement $\gamma_{\bar{w}}$. A. 5-mer motifs with strong central nucleotide (C/G), $r^2 = 54\%$. B. 5-mer motifs with weak central nucleotide (A/T), $r^2 = 7\%$

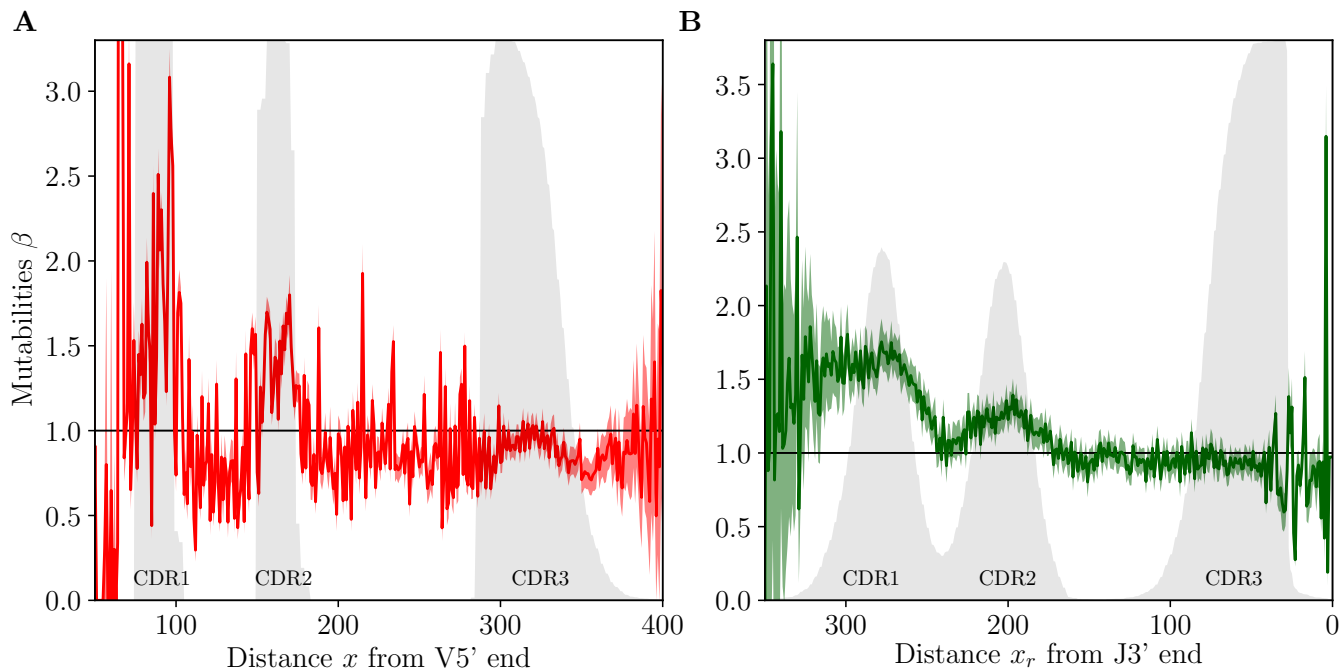

FIG. S3: Alternative position definition. We compare the models based on different position definitions: x, the distance from the $5'$ end of the V gene (A, red) and x_r, distance from the $3'$ end of the J gene (B, green). Error bars correspond to 95% confidence intervals. Frequency at which a given position belongs to a CDR region is indicated with the grey shaded areas.

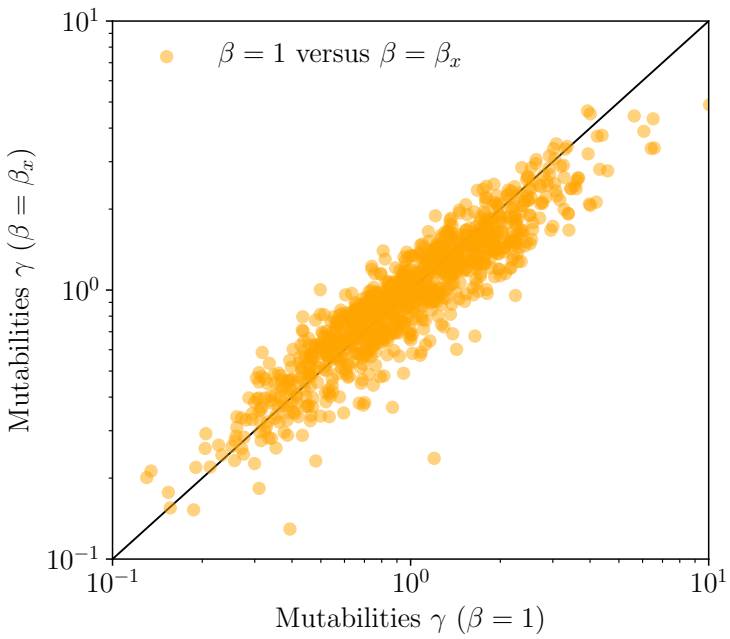

FIG. S4: Introducing explicit position dependence $\beta \neq 1$ influences the context-dependent rates. We compare the γ mutabilities from the full model with the parameters of the purely context-dependent model.

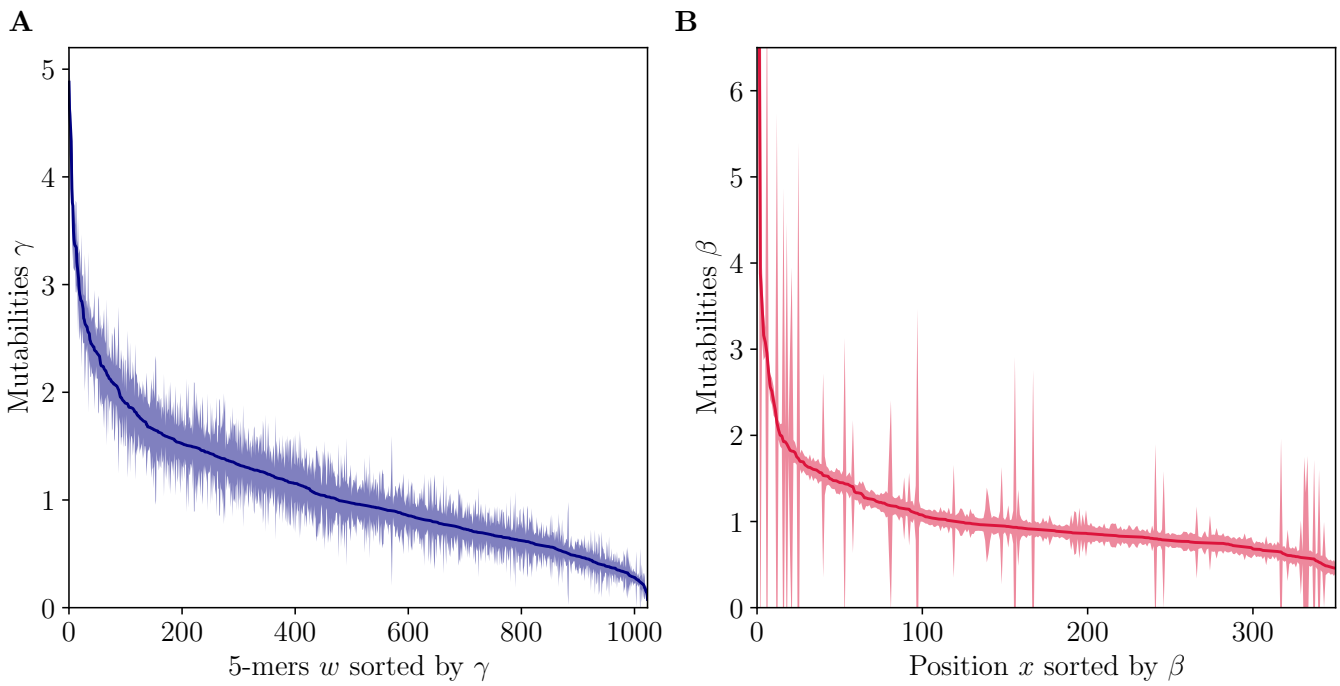

FIG. S5: Analysis of the uncertainty of estimate parameters γ (A) and β (B). The shaded area indicates the 95% confidence interval envelope. Motifs and positions are sorted by their respective mutabilities

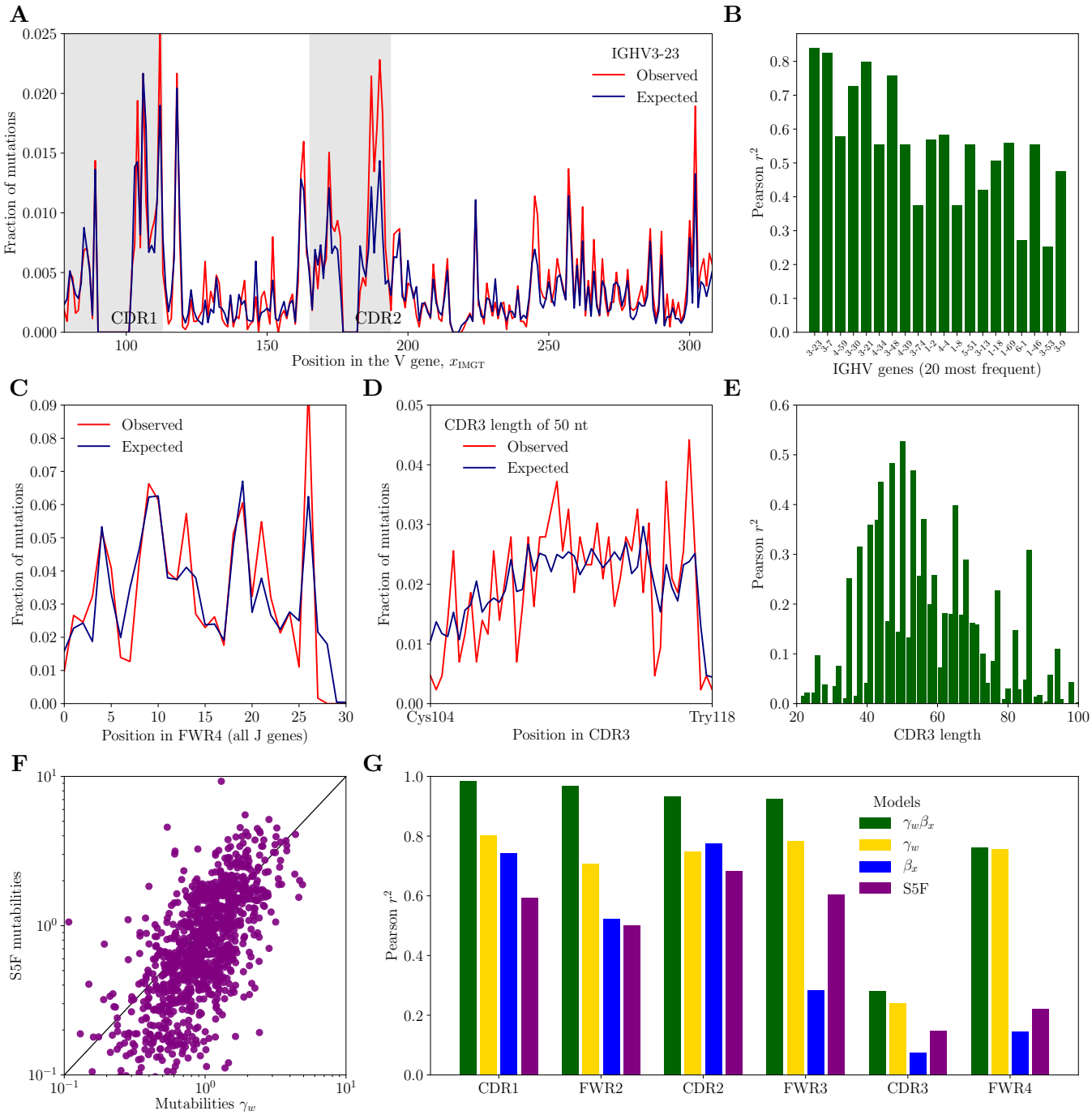

FIG. S6: Model performance when model was trained on 2/3 of data, and tested on the remaining 1/3. A. Example mutation profile in the most common V gene. B. Model performance across V genes. C. Mutation profile in the FWR4 region. D. Example mutation profile in the CDR3 region for CDR3 length of 50 nts. E. Model performance across CDR3 lengths. F. Comparison with the S5F model. G. Summary of models performance across sequence regions.

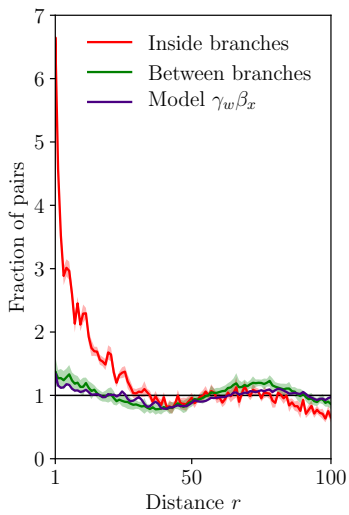

FIG. S7: Normalized fraction of pairs $n(r)$ of mutations encoded in the same tree branches (red) and pairs of mutations between different branches (green). Shaded area corresponds to 95% confidence interval.

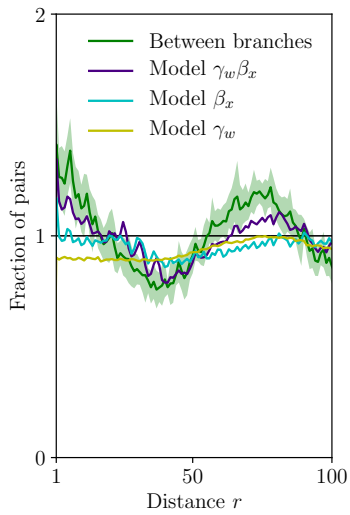

FIG. S8: Normalized fraction $n(r)$ of pairs of mutations between different branches (green). Context- and position-dependent models prediction $n_m(r)$ for this quantity are presented for the full ($\mu_{s,x} = \gamma_w\beta_x$), purely context-dependent ($\mu_{s,x} = \gamma_w$) and purely position-dependent ($\mu_{s,x} = \beta_x$) models. Shaded area corresponds to 95% confidence interval.