Dynamics of memory B cells and plasmablasts in healthy individuals

Andrea Mazzolini,[1,2,3] Aleksandra M. Walczak,[1,*] and Thierry Mora[1,*]

[1] *Laboratoire de physique de l'École normale supérieure, CNRS, PSL University, Sorbonne Université, and Université Paris-Cité, 75005 Paris, France*
[2] *Department of Physics and INFN, University of Turin, 10125 Turin, Italy*
[3] *INFN, Sezione di Bari, Italy*

Our adaptive immune system relies on the persistence over long times of a diverse set of antigen-experienced B cells to encode our memories of past infections and to protect us against future ones. While longitudinal repertoire sequencing promises to track the long-term dynamics of many B cell clones simultaneously, sampling and experimental noise make it hard to draw reliable quantitative conclusions. Leveraging statistical inference, we infer the dynamics of memory B cell clonal dynamics and conversion to plasmablasts, which includes clone creation, degradation, abundance fluctuations, and differentiation. We find that memory B cell clones degrade slowly, with a half-life of 10 years. Based on the inferred parameters, we predict that it takes about 50 years to renew 50% of the repertoire, with most observed clones surviving for a lifetime. We infer that, on average, 1 out of 100 memory B cells differentiates into a plasmablast each year, more than expected from purely antigen-stimulated differentiation, and that plasmablast clones degrade with a half-life of about one year in the absence of memory imports. Our method is general and could be applied to other longitudinal repertoire sequencing B cell subsets.

I. INTRODUCTION

The adaptive immune system has to recognize a practically infinite reservoir of antigens derived from pathogens. This is achieved thanks to the huge diversity of lymphocyte receptors created through V(D)J recombination, which randomly generates new sequences [1, 2], and can lead to $10^8 - 10^{10}$ distinct receptor types in each organism [3, 4]. This diversity is further increased in B cells through affinity maturation [5, 6], after rounds of somatic hypermutations and selection against encountered antigens [7, 8]. This process releases cells in the blood that are evolutionarily related, forming a clonal family specifically tuned for the antigen against which it evolved. These cells divide into memory cells and plasma cells [9] whose role is to keep immunological memory against encountered antigens. In particular, memory cells can differentiate into antibody-secreting plasmablasts upon antigen re-encounter [10], providing a quick and effective response to reinfections.

High-throughput DNA-sequencing experiments have significantly increased our knowledge about the immune repertoire diversity [7, 11–15] and evolutionary properties of B-cell evolution [16–20]. Furthermore, sequencing experiments taken at multiple time points opened up several questions related to B-cell repertoire dynamics [17, 21–23] and, in particular, about the stability of the memory cell repertoire and its relation with plasmablasts in healthy conditions [24, 25]. The memory repertoire was found to be relatively stable at the level of comparing statistics of common clones between time points and across experimental replicates at the same time. The authors in [24] hypothesized that memory-to-plasmablast differentiation does not happen only upon specific antigen re-encounter [26], but also through cross-reactivity with different persisting antigens or antigen-independent mechanisms [27, 28]. However, the timescales of these events remain unknown.

Learning about the dynamics of repertoires from data requires a lot of care. Limited sampling and noise in repertoire sequencing experiments leads to low clonal overlap between samples, including between technical replicates taken at the same timepoint [24]. Such observations in longitudinal studies may wrongly suggest a large turnover. In this work, we aim to provide a quantitative description of the repertoires of two cell types: memory B cells and plasmablasts. To this end, we apply statistical inference techniques to sequencing data from the blood of healthy individuals. To disentangle the actual repertoire dynamics from the large sampling and experimental noise that confounds it, we adapt a previously introduced method used in the context of T-cell repertoire [29–31]. Our inference provides biologically informative, previously unknown quantities, such as the half-life of memory clones, how quickly the repertoire replaces its B-cell clones, or the average differentiation rate of memory cells into plasmablasts. The model accurately describes the statistics of clone abundances across longitudinal experiments, providing probabilistic predictions of future cell counts given present observations in healthy conditions. This can provide a useful baseline for tracking the evolution of immunity and response to chronic and recurring pathologies over time.

*Corresponding authors. These authors contributed equally.

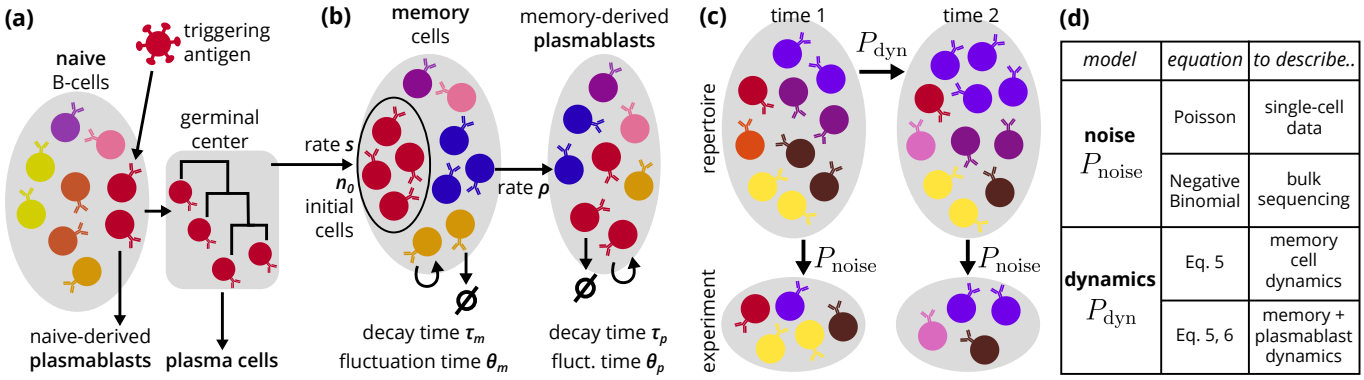


FIG. 1: **Modeling and inference scheme.** (a): Simplified schematic of biological processes that lead to the creation of memory-cell clones starting from the naive repertoire. (b): Processes involving memory and plasmablast clones that are included in the studied dynamical models (see Sec. II A). Empty sets represent cell death and self-arrows represent proliferation. (c): Inference scheme of the repertoire clone dynamics between two time points. The variation of frequencies in the body is modeled with $P_{\rm dyn}$. Experiments provide counts that convey noisy estimates of the real abundances through the noise model $P_{\rm noise}$. (d): Different choices for the noise and the dynamical model discussed in the paper.

II. RESULTS

A. Scheme for inferring B cell population dynamics from noisy data

Our goal is to infer a probabilistic dynamical model, $P_{\rm dyn}$, to describe the evolution of the size of B cell clones over long times. We focus on two cell types: memory B cells and memory-derived plasmablasts. A summary of our model assumptions informed by their known biology is illustrated by Fig. 1a,b. Memory B cell clones emerge from affinity maturation, with mean rate s and average initial size n_0. They then follow stochastic dynamics subject to random environmental fluctuations. Mathematically, their clone sizes are described by a Geometric Brownian Motion (GBM), following [31, 32], see Methods. In that model, the size of each clone is subject to random fold-changes over a typical timescale θ_m, which we refer to as the fluctuation time. Overall, these changes, which can be caused by cell proliferation, death and differentiation, tend to decrease the size of clones with characteristic time τ_m, called decay time. We chose that model because it is the simplest possible model for stochastic growth and decay, but also because it predicts power-law distributions of clone sizes [32], consistent with experimental observations [33, 34]. Within this framework, the exponent of the power law of the cumulative size distribution, α, is linked to the dynamical parameters θ_m and τ_m through the relation $\alpha = 2\theta_m/\tau_m$.

Memory cells can differentiate into plasmablasts with rate ρ, providing a rapid response against pathogens encountered in previous infections. After differentiation, plasmablast dynamics is also subject to environmental fluctuations and decay, which can be modeled by a GBM with timescales θ_p and τ_p, similarly to memory cells. The model we use to describe the dynamics of plasmablast repertoires combines the contributions of the GBM with a source of new cells from memory B cell differentiation (see Methods).

The biological processes we aim to describe are discrete, heterogeneous in time and across clones—for example, infections cause a small subset of clones to undergo quick antigen-driven bursts. The applicability and interpretation of our model, which summarizes these heterogeneities into stochastic terms, make sense only when looking at samples that are well separated in time, on timescales larger than those typical events of a few weeks. Model rates represent a net average of the changes occurring over those long times. For example, ρ should be interpreted as the fraction of memory cells that differentiate into plasmablasts in a year, rather than the rate of each individual cell.

To infer those dynamical models, we use longitudinal repertoire sequencing data. Repertoire sequencing only probes an infinitesimal fraction of the immune repertoire, yielding a noisy estimate of clonal frequencies. At the same time, the repertoire dynamics themselves are expected to be stochastic and modeled by $P_{\rm dyn}$, following, for example, the GBM described above. To disentangle these two sources of stochasticity — experimental noise and biological fluctuations — we developed a statistical inference scheme based on [29–31] and presented in the Methods section, with additional detail in the SM. It relies on two probabilistic models (Fig. 1c): the dynamical model $P_{\rm dyn}$ and a noise model $P_{\rm noise}$, which describes how actual clone sizes translate into observed sequence counts. It is chosen to be a simple Poisson sampling model or a more complex Negative Binomial model, depending on the nature of the data (more on this below). The different types of dynamical and noise models used in this paper are summarized in Fig. 1d.

B. Longitudinal, replicated B cell repertoire datasets

The method takes as input longitudinal datasets of memory B cells and circulating plasmablasts, and must also include technical replicates to calibrate the noise model. We collected two open-access longitudinal datasets from human Peripheral Blood Mononuclear Cells (PBMC) that fit these requirements [24], [25].

The first dataset [24], which we call *Dataset A*, comprises repertoires from two healthy individuals, aged 50 and 69, over a span of 10 and 6 years respectively, with up to 5 replicates per time point. All repertoires were obtained by single-cell sequencing, which allows for pairing heavy and light chains, as well as phenotyping. The samples, including their cell counts, are summarized in Fig. S1.

The second dataset [25], *Dataset B*, includes IGH (heavy-chain) repertoires from 6 healthy individuals, aged 23 to 39, spanning a period of 1 year, with up to 2 replicates per timepoint (see details and sequence counts in Fig. S2). Cells were first sorted into memory B cells, plasma cells, and plasmablasts, and their repertoire was sequenced in bulk from mRNA, using Unique Molecular Identifiers to mitigate PCR noise and biases. We only used the repertoires of 3 of the 6 individuals (AT of age 23, IM of age 39, and IZ of age 33) for whom samples were sequenced deeply enough for all time points and cell types.

Data was processed using standard tools. We define a clone or clonal group as the set of B cells that descend from the same initial V(D)J recombination event. Clonal groups of sequences were inferred using HILARy [35] (see Methods for details).

C. Distinct noise profiles of single-cell versus bulk repertoire sequencing

The first step of the inference scheme is to learn a model to quantify experimental and sampling noise (P_{noise} in Fig. 1). The noise model comes in three versions of increasing complexity and generality (see Methods): the Poisson Model (zero parameters), a Negative Binomial Model with a linear relation between variance and mean (1 parameter), and a Negative Binomial Model with a power-law dependence of variance and mean (2 parameters). In addition, all 3 models assume that the distribution of clone frequencies f follows a power law, parametrized by 2 additional parameters: its exponent α and the minimal clone frequency $f_{\min}$. The form of the clone size distribution is not strictly part of the noise model and is not expected to impact its inference, but it is necessary as a prior for the fitting procedure. We chose its form according to empirical observations.

To determine which model to use, we performed the inference of these three models on all the timepoints for which at least 2 replicates are available and compared

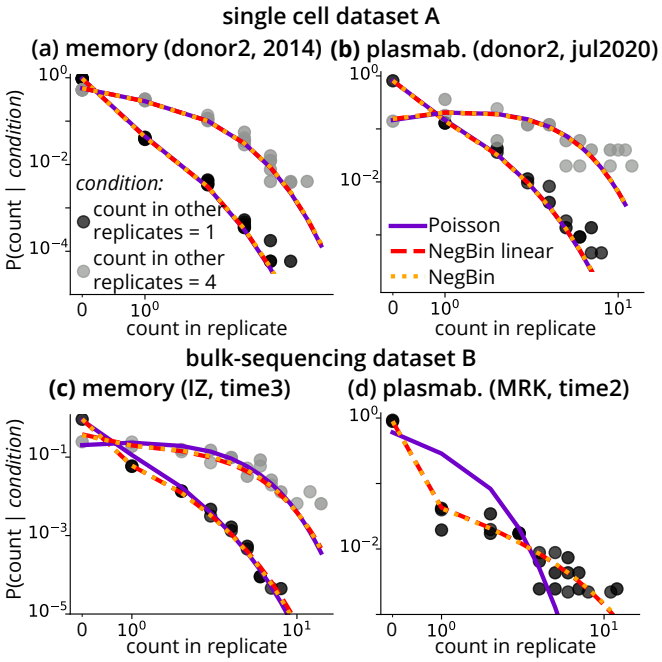


FIG. 2: **Validation of the noise model.** Shown is the conditional distribution of counts of a clone in one of the replicates, given the total count of the same clone in all other replicates (1 in black, 4 in gray). Each replicate is downsampled in such a way that every replicate has the same size to compare them with the same theoretical prediction (which depends on the sample size). Each panel shows a different experiment, corresponding to one donor in each of *Dataset A* (panels a and b), which uses single-cell sequencing, and *Dataset B*, which uses mRNA bulk sequencing (c and d), and to memory (a and c) as well as plasmablast (b and d) subsets. In (d) we do not report the conditioning on 4 because the points partially overlap with the other case. For each sample, predictions from the inferred noise models are shown for the Poisson model, the best Negative Binomial Model with a linear relation between variance and mean (NegBin linear), and the best Negative Binomial Model with a power relation between variance and mean (NegBin).

their log-likelihood to assess their ability to accurately describe the data (Fig. S3 and S4). For the single-cell *Dataset A*, all three models performed equally well, leading us to choose the simpler Poisson Model for the rest of the analysis. This is in agreement with previous observations in single-cell RNAseq data [36, 37], consistent with a model in which the main source of noise is due to sampling single cells. For *Dataset B*, which was obtained by bulk sequencing of mRNA, we found that the Negative Binomial Models performed slightly better than Poisson in the memory-cell samples and significantly better in plasmablasts. However, the more general model with a power-law dependence between variance and mean did not offer a substantial improvement over the model with a linear one. In addition, the power-law model was harder to optimize, finding parameters at the boundaries of the allowed space and violating prior knowledge, notably on

the predicted power-law exponent of the clone size distribution (Fig. S4b). Therefore, for *Dataset B* we used the linear Negative Binomial Model, with a distinct linear coefficient inferred for each timepoint. The overdispersion of the measured count relative to mere sampling (Poisson) noise can be explained by the additional expression, reverse-transcription, and amplification noises caused by the mRNA bulk sequencing protocol [29–31].

To test the noise model, we used it to predict the distribution of counts of a clone in one of the replicates, conditioned on its total count across the other replicates. Comparing this prediction to conditional histograms of clone sizes collected over the entire repertoire from a single sample yields excellent agreement (Fig. 2), except for the Poisson Model applied to *Dataset B*, especially for plasmablasts (Fig. 2d), consistent with our log-likelihood performance assessment. Finally, the distribution of aggregated counts across all replicates follows a power law that is well predicted by the model, which validates the choice of the frequency distribution $\rho(f)$ (Fig. S5).

The inferred noise model should not depend on the size and number of used replicates. To check for this robustness, we repeated the noise inference on the 2014 memory sample of donor 2 of *Dataset A* using various subsets of the 5 replicates (Fig. S3e). The number of replicates mostly affected the inference of the power law exponent and minimal frequency, but only weakly, with a still negligible gap between the Negative Binomial and Poisson likelihoods, confirming our choice to use the Poisson model for this dataset.

D. Memory B cell clones persist for decades

We next applied our inference method to learn the dynamical parameters of the memory repertoire: the fluctuation time, θ_m, the net decay time, τ_m, the rate of introduction of new memory clones, s, and their initial size, n_0. Parameter optimization was performed using Monte Carlo integration, as explained in detail in Sec. S3 and illustrated in Fig. S6. The results are summarized in Table I. The net decay time of individual clones in the memory repertoire of individuals in *Dataset A* is ≈ 15 years, or a half-life of $\ln(2) \times 15 \approx 10$ years, meaning that, on average, clones reduce their size by half in 10 years. The noise timescale is ≈ 10 years, meaning that it takes less than a decade for individual clone fluctuations to lead to a two-fold increase or decrease.

Inference on *Dataset B* does not allow us to learn a precise number for τ_m, but puts on it a lower bound of 4 years (Sec. S3F and Fig. S7), which is consistent with the estimate of 15 years from *Dataset A*. The reason for that uncertainty is that the time interval between samples is only 1 year. Within that short interval, the data do not show evidence of a decrease in time, and the inference cannot distinguish between scenarios with a τ_m larger than the bound.

To assess the model's validity, we perform a similar test

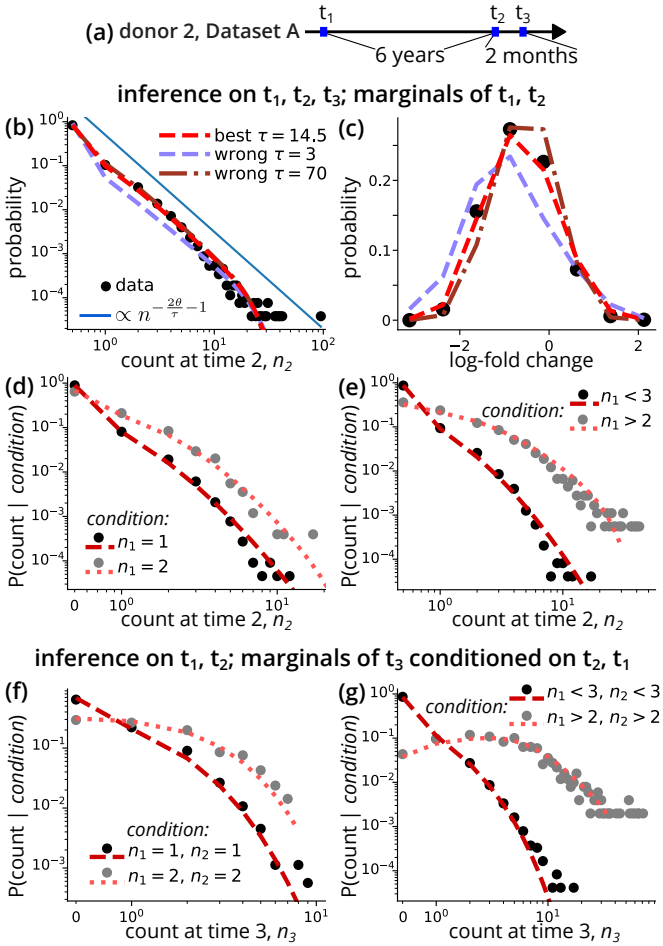


FIG. 3: **Validation of the Geometric Brownian Motion model for memory cells.** Statistics of memory B-cell counts and its prediction by the GBM model in donor 2 from *Dataset A*. The model prediction is shown for the best fit value of $\tau_m = 14.5$ yr and $\theta_m = 10.1$ yr, as well as for $\tau_m = 5$ and 70 yr, keeping the ratio θ_m/τ_m constant to its best-fit value. (a): Summary of the samples and timepoints of donor 2. (b): Clone size distribution at the second time point. (c): Distribution of log-fold changes in clone counts between t_1 and t_2. We only include clones with count $n_1 > 2$. (d), (e): Counts at time 2 given a condition on the counts observed at time 1. Black and gray points refer to two different conditions. (f),(g): Similar to (d), (e) but for the third time point. To make a fully independent prediction, for these two panels the model parameters were re-learned using only the first two time points.

to the one performed for the noise model. We predict the distribution of clone counts, and of their fold change, as a function of their count at an earlier time point (instead of a replicate, as we did for the noise model). Despite its simplicity, the model fits all the distributions very well (Fig. 3b-g). To show that the goodness of fit depends sensitively on learning the correct parameters, we also compared the prediction of sub-optimal models where the values of τ_m and θ_m were varied while keeping their ratio fixed to the inferred value. Theory predicts that

		Noise model	**Memory cell dynamics**					
	age (year)	Neg. Bin. parameter, a	decay time, τ_m (year)	fluctuation time θ_m (year)	power law exponent, α	initial clone size, n_0	rate of new clones, s (1/year)	number of clones, N_{clone}
donor 1 (A)	69	0	14.6 ± 1.3	9.3 ± 0.6	1.28 ± 0.05	$4 \cdot 10^4$–$4 \cdot 10^5$	$(3.6 \pm 0.6) \cdot 10^3$	$(5.5$–$6.8) \cdot 10^5$
donor 2 (A)	50	0	14.5 ± 1.3	8.7 ± 0.7	1.20 ± 0.04	$1 \cdot 10^4$–$1 \cdot 10^5$	$(9.0 \pm 1.6) \cdot 10^3$	$(1.2$–$1.5) \cdot 10^6$
AT (B)	23	0.086 ± 0.005	> 3.2	> 2.1	1.29 ± 0.01	$5 \cdot 10^3$–$5 \cdot 10^4$	$< 1.5 \cdot 10^5$	$(4.0$–$5.1) \cdot 10^6$
IM (B)	39	0.059 ± 0.004	> 4.0	> 2.7	1.30 ± 0.01	$6 \cdot 10^3$–$6 \cdot 10^4$	$< 9.4 \cdot 10^4$	$(3.4$–$4.3) \cdot 10^6$
IZ (B)	33	0.167 ± 0.004	> 4.8	> 3.3	1.35 ± 0.01	$3 \cdot 10^3$–$3 \cdot 10^4$	$< 1.8 \cdot 10^5$	$(6.8$–$8.8) \cdot 10^6$

TABLE I: **Inferred parameters of the Geometric Brownian Motion for memory B-cell dynamics.** The third column corresponds to the average parameter a of the Negative Binomial model (see Methods) and its average error across the time points (see Fig. S4a for values at each single time point). The other columns give the GBM parameters (τ_m, θ_m, s, and n_0), and their derivatives, including the predicted exponent $\alpha = 2\theta/\tau$, and the repertoire diversity N_{clone}. For n_0 and N_{clone} we show a range of values corresponding to an assumed total number of memory B-cells $N_{\text{cell}} = 10^{10}$–$10^{11}$. For details about inference and error estimation see Sec. S3.

the clone size distribution should follow approximately a power law, whose exponent depends only on that ratio, predicted to be $\alpha = 2\theta_m/\tau_m \approx 1.2$-$1.3$. Consistently, that distribution is well captured even by the sub-optimal models (Fig. 3b). By contrast, the distribution of log fold changes is only well reproduced by the inferred model ($\tau_m = 14.5$ yr), and not by the sub-optimal ones ($\tau_m = 3$ and 70 yr). A shorter time scale ($\tau_m = 3$ yr) predicts a distribution that is more shifted towards negative values, indicating faster turnover and hence larger decay, while a larger $\tau_m = 70$ yr predicts smaller decay than the data (Fig. 3c). The inferred model can correctly predict the full distribution of clonal counts at a timepoint t_2, given its abundance at an earlier timepoint t_1 (Fig. 3d-e). The model generalizes well to held-out samples (i.e. not used for inference) of the same individual: the distribution of counts on a third time point conditioned on counts on the first two shows excellent agreement with the data (Fig. 3f-g). The model inferred on donor 2 of *Dataset A* is also successful in predicting the marginals of donor 1 (Fig. S8a), consistent with the fact that their inferred timescales are very similar (Table I). This gives us another test of robustness and suggests that the inferred dynamical models generalize across people.

A characteristic decay time of $\tau_m \approx 15$ years implies that clones can, in fact, survive for much longer than that. To see this, we can use the model to compute the probability of survival of a clone (i.e. that there remains at least one cell from that clone in the organism) over time, as a function of its cell count in a given repertoire sequencing experiment. Strikingly, after 25 years, even clones represented by just a single cell in the experiment ($n = 1$) are predicted to survive almost certainly, while clones with larger counts will very likely never go extinct in our lifetime (Fig. 4 a). The reason for this long survival, despite a more modest decay time, is that experiments only capture a very small fraction of all memory B cells (10^5 out of $\approx 10^{10}$–10^{11}). Even clones with a single cell in the sampled repertoire are typically abundant in the whole repertoire (average count $\approx 10^4$–10^5, using the parameters of Fig. 4). Even with decay, they are likely

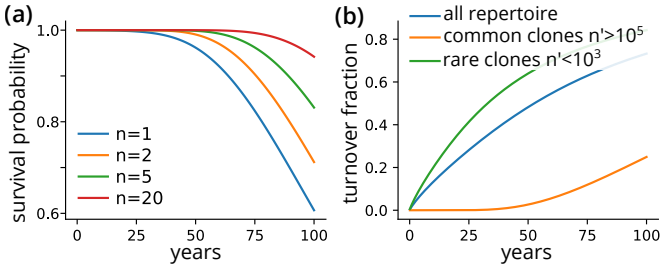


FIG. 4: **Probability of future survival/turnover.** (a): Survival probability of a clone after t years given an experimental observation n, Eq. S11. We have used $\tau = 14.5$, $\theta = 9$, $N_{cell} = 10^{10}$, $n_0 = 10^4$, and $M = 10^6$. (b): Fraction of clones that go extinct after t years (Eq. S10 of Supplementary Material), computed with the same parameters as before. The three lines consider clones in different ranges of abundances as specified by the legend.

to survive. The framework also allows us to compute the probability of observing a clone in future experiments as a function of its count in the current one (Fig. S8b).

We can use the model to answer a related question: after t years, what is the fraction of clones that will go extinct and be replaced by newly generated ones? Fig. 4b shows that repertoire renewal is quite slow, with replacement of half the clones after 50 years (blue line for the whole repertoire). The turnover depends on clone abundances: large clones (with total number of cells in the organism $> 10^6$, which are likely sampled in an experiment, show negligible turnover before 50 years. Even taking the lowest estimate of $\tau_m = 5$ yr from *Dataset B*, half of the repertoire would be renewed after around 25 years (Fig. S8c). Overall, these estimates suggest that the memory B cell repertoire is stable over decades.

E. Rate and size of new memory B cell clones, and their total diversity

In addition to τ_m and θ_m, we inferred two additional parameters of the Geometric Brownian Motion: s, the rate of introduction of a new memory B cell clone, and n_0, the initial number of cells in each new clone. The rate s provides an upper bound on the rate with which new germinal centers (GC) appear, since each GC may produce multiple clones. However, if we assume that all cells from a GC belong to the same clone [6], then s may be directly interpreted as the rate of appearance of new productive GCs.

We found that the absolute initial memory clone size n_0 was underdetermined under our statistical approach. However, the ratio of initial clone size to the total number of memory cells, n_0/N_{cell}, is well inferred (Fig. S6b). We can thus infer n_0 by using estimates of the total number of memory B cells $N_{\text{cell}} \approx 10^{10}$–$10^{11}$ [3, 38, 39], for which the GBM model makes a mathematical prediction as a function of the 4 model parameters (τ_m, θ_m, s, n_0), see Sec. S3. That additional constraint allows us to obtain n_0 as a function of the other 3 parameters, which we inferred reliably. The results of this procedure, reported in Table I, lead to an estimated $n_0 \sim 10^4$ memory cells per new clone produced by a germinal center. Recall that n_0 is an average, as the output of germinal centers is known to be highly variable [40]. Note that n_0 implies an initial frequency of new clones of less than one part in a million ($10^4/10^{10} = 10^{-6}$), meaning that new clones are typically not expected to be found by a repertoire sequencing experiment. This explains why n_0 is hard to infer from the data.

The rate of appearance of new memory clones, s, is estimated to be 10^3–10^4 per year. Assuming that each GC produces only one or a few clones, that implies thousands of GC formed every year. These estimates are obtained for *Dataset A* only, as *Dataset B* only allows us to put an upper bound on it.

By deriving exact analytical formulas from the model (see SI Sec. S3A), we can estimate the real diversity N_{clone} of clones, i.e. the total number of memory B cell clones in the organism. It can be expressed as a mathematical function of the 4 parameters, yielding an estimate of $N_{\text{clone}} \approx 10^6$.

Interestingly, our data on 5 individuals suggests that the ratio of initial clone size to the total number of memory cells n_0/N_{cell} increases with age, and that the total diversity N_{clone} decreases as a result (Fig. S9). This would mean that younger individuals either have a larger pool of memory cells (higher N_{cell}), or that they produce fewer memory cells from each germinal center, resulting in higher diversity. This last observation would be in agreement with previous work on age-dependent diversity of B-cells [41]. However, these conclusions are based on only 5 points, and there can be dataset-dependent biases, requiring confirmation from future analysis of other datasets.

	differentiation rate ρ (1/year)	decay time τ_p (year)	fluctuation time θ_p (year)
donor 1 (A)	0.003–0.03	0.5–3	0.1–0.3
donor 2 (A)	0.01–0.1	0.25–1	0.05–0.1

TABLE II: **Inferred parameters of the plasmablast-memory coupled dynamics.** The parameters are learnable only for *Dataset A*. The inference uses the memory parameters learned at the previous step, in Tab. I. Each parameter shows a range that depends on the unknown variability of the ratio of plasmablasts to memory cells, $r \sim 0.01$–0.1, and the fraction of plasmablasts that differentiate from memory cells with respect to the total plasmablasts, $f \sim 0.3$–0.8.

.

F. 1% of memory B cells differentiate into plasmablasts every year

We next applied our inference approach to the plasmablast repertoires. We only managed to successfully perform that inference for *Dataset A*, likely because stronger noise in *Dataset B* impeded learning (see Sec. S4 and Fig. S10 for details). The model now has three learnable parameters: the decay time τ_p, the fluctuation time θ_p, and the rate of imports from the memory pool ρ (see above and Methods). Because repertoire sequencing only allows us to capture relative frequencies and not absolute numbers, the inference scheme leaves two unidentifiable quantities by construction (see Methods): the fraction of plasmablasts that originate from memory cells, ϕ, and the ratio between the number of plasmablasts and memory cells, r. We observe that around 25-30% of sampled plasmablasts can be assigned to clones that are represented in the sampled memory pool, setting a lower bound $\phi \geq 30\%$. Because the sampling of the memory pool is very limited, the real value of ϕ is likely much larger. To account for this uncertainty, we performed the inference fixing ϕ to values ranging from 30% to 80%. The ratio r of plasmablast to memory cells depends on many factors, and probably varies in time, but rough estimates from the literature [42] as well as the ratio of plasmablast to memory in experiments suggest a value ranging from $r = 1\%$ to $r = 10\%$. Results reported in Table II show a range that is a consequence of this variability and the error associated with the maximum likelihood procedure. The inferred values of the 3 parameters (ρ, τ_p, θ_p) while fixing ϕ and r to various values are reported in Fig. S11.

The memory-to-plasmablast differentiation rate is found to be around 0.01/year or larger: on average, at least 1 out of 100 memory cells differentiates into a plasmablast each year. The net decay time of plasmablast clones ranges from a few months to a few years, which is 10 to 100 times faster than the 15 years of the memory B cell clones. Note that this decay time is still large compared with the life span of plasmablasts, which is a few days [43]. This implies that clone sizes must be main-

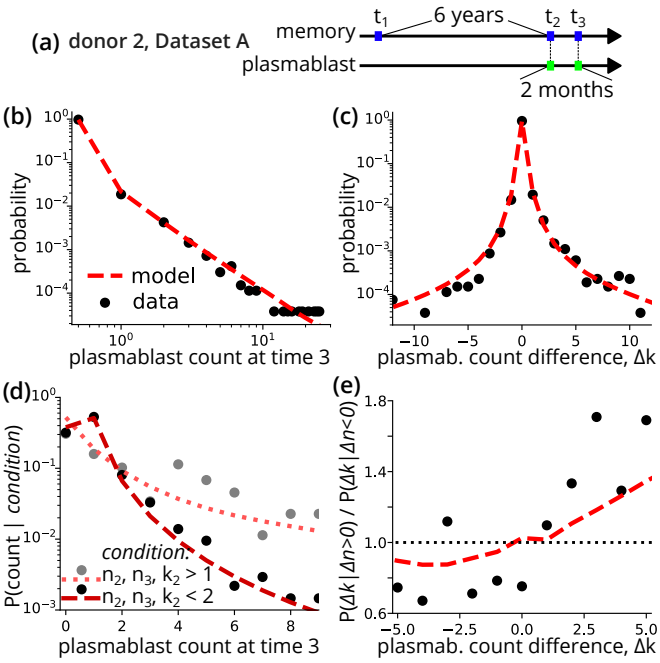


FIG. 5: **Validation of the plasmablast-memory coupled dynamics.** Conditional distributions of plasmablast counts from donor 2 from *Dataset A*. The model uses the best-fit parameters inferred for $r = 0.02$ and $\phi = 0.5$, which are $\tau_p = 0.4$ and $\theta_p = 0.06$ years, implying a differentiation rate $\rho = 0.026$/year. (a) Summary of the samples and timepoints of donor 2. (b) Plasmablast clone size distribution at the third time point, conditioned on the clone being present in the memory sample at the first time point. (c) Distribution of the count difference Δk between t_2 and t_3. (d) Distribution of plasmablast counts at t_3 given a condition on the second and third memory-sample counts (labeled by n) and on the second plasmablast-sample counts (labeled by k). Black and gray points refer to the two different conditions. For the second condition (black points), we select only plasmablasts with clones also present in at least one of the memory samples, i.e., we select only the plasmablasts that are memory-derived (this is automatically satisfied for the first condition). (e) Odds ratio of two probability distributions over the plasmablast count difference Δk. The numerator is conditioned on an increase in the corresponding memory counts, $\Delta n = n_3 - n_2 > 0$, and the second to decrease $\Delta n < 0$. A ratio larger than 1 means that clones with a given Δk are more likely to be observed when the memory counts grow.

tained by cell proliferation to counter cell death, consistent with reported proliferation activity in plasmablasts [44]. The fluctuation time θ_p is small compared to the decay time τ_p, meaning that fluctuations in that proliferation activity are large. These large variations may originate from clonal heterogeneity in the cell differentiation, division, and death rates, which is not explicitly described by our model but would effectively lead to large fluctuations.

We validated the model by comparing its predictions to data, focusing on the model inferred from *Dataset A*'s second donor (Fig. 5a). The plasmablast clone size distribution (Fig. 5b) is well reproduced, as well as the distribution of fold changes over 2 months (Fig. 5c). Because both τ_p and θ_p are short relative to the memory parameters, the turnover dynamics of plasmablasts is fast, meaning that the plasmablast repertoire is a slightly delayed mirror image of the memory one. Hence, the power law observed in its clone size distribution is largely inherited from that of the memory pool.

In addition, the model correctly predicts the distribution of plasmablast clone sizes at the third time point given their recorded abundance in other samples (Fig. 5d), including the memory pool. Lastly, the model makes a specific prediction about how the change of clone size in the plasmablast pool is influenced by the change of clone size in the memory pool of the same clone, which agrees with the data (Fig. 5e): clones that grew in the memory pool are more likely to grow in the plasmablast pool in the same interval ($p = 3.6\ 10^{-9}$, Mann-Whitney U test).

G. Robustness of the inference

A limitation of our approach is its dependence on the choice of software for reconstructing clonal families. To check for robustness to that choice, we repeated the analysis using the alternative ChangeO tool instead of HILARy (see Methods) for donor 2 of *Dataset A*. Inferring the noise model on these families yielded similar parameters (Fig. S12a). The likelihood of the Poisson and the linear Negative Binomial models was almost the same (Fig. S12b), justifying the use of the Poisson model for downstream analysis. Similarly, the parameters of the memory and plasmablast models inferred with the ChangeO families showed negligible differences with HILARy (Table S1). This indicates that our inferred parameters are insensitive to the choice of clustering algorithm.

In general, clustering algorithms can wrongly merge together distinct clones and thus slightly overestimate their decay time τ_m. To control for this effect, we repeated the analysis by defining a "clone" as a set of identical heavy-chain sequences. Under that definition, any hypermutation creates a new clone. The inferred decay time τ_m can thus be seen as a lower bound. Applying inference with that definition indeed yields a smaller τ_m than with the HILARy clones, but still large (~ 10 years, see Table S2), confirming our main conclusions.

Finally, the analysis relies on the FACS sorting that can wrongly sort some B cells between memory cells and plasmablasts. To assess the impact of mis-sorting, we artificially replaced a fraction of memory cells by plasmablasts collected at the same time point and repeated the inference of the GBM (Fig. S13). As expected, the inferred average decay time τ_m decreases as the fraction of plasmablasts in the sample increases, since plasmablast dynamics are faster. However, this change is only visible when a large fraction (more than 10%) of cells are

mis-sorted, making our conclusions also robust to that potential source of errors.

III. DISCUSSION

We presented a framework to quantify the dynamical changes of memory B-cell and memory-derived plasmablast repertoire in healthy donors. Our approach infers the parameters of stochastic dynamics models while accounting for experimental noise. The results give us access to biologically informative quantities and provide an accurate statistical description of clone counts in longitudinal repertoire sequencing experiments.

The method is, however, reliant on the quality and structure of the dataset. It requires experimental replicates to infer a noise model, which is particular to each protocol and laboratory setting. We found that two distinct models, Poisson and Negative Binomial, best explained the 2 datasets we analyzed. Another requisite is that longitudinal samples be properly spaced. For example, the one-year time window of *Dataset B* was not enough to learn the longer degradation timescale of memory cells. Also, experimental noise and sequencing depth are important: the large negative binomial noise of *Dataset B* concealed the signal of the plasmablast dynamics, preventing inference in that case.

By modeling the memory cell dynamics with a Geometric Brownian Motion, we accurately captured the statistics of clone counts in samples, including in future ones not used for inference. The inferred parameters teach us that the half-time of memory cell clones is around 10 years. This implies that clones that are observed in experiments survive for tens of years. This number differs from existing measurements on the lifetime and turnover of lymphocytes [45–48], which instead describe the life expectancy of *single cells*. This long lifespan of clones arguably correlates with immune memory and is consistent with the observed time scale of protection against encountered pathogens [49, 50]. A previous analysis of T-cell clonal lifespans led to estimates of the same order of magnitude, from several years to decades [31], and increased with age, consistent with the reduction of thymic output. Because of the limited number of individuals, our analysis did not allow us to probe the age dependence of memory B cell clonal survival. Because the half-life of 10 years is much larger than the lifespan of a single memory cell, which is only a few months [48, 51], our estimates imply that the maintenance of memory clones must involve cell proliferation to compensate for cell death, rather than through the existence of long-lived quiescent cells [10, 52]. In that picture, individual cells divide and die quickly, with the two effects balancing each other and keeping the clone size stable for years. As a result, repertoire turnover, i.e. the replacement of old memory clones by new ones, is relatively slow and mostly affects small clones that are not sampled in experiments.

The inference also gives us access to parameters connected with the emergence of memory clones and, therefore, with properties of germinal centers. We inferred that every year around 1,000-10,000 new memory clones appear, each one of them producing $10,000$ memory B cells, and that there are around $N \sim 10^6$ different memory clones in our body. It is difficult to compare these numbers to direct estimates from the literature. However, we can use the parameters used in standard models of affinity maturation [53] to get a rough estimate of germinal center output. Starting from a few thousand cells in a germinal center, after two rounds of replication twice a day, a few tens of thousands of cells are tested against the antigen each day. Of those, between 10% and 33% pass the selection step, which then have around a 5% probability of differentiating into memory cells, making a few hundred new cells per day. After a few weeks, the memory cell output would be around 10^4, consistent with our estimate. These GC-related numbers, however, should be taken with a grain of salt: we expect them to mostly affect the statistics of small clones, which by definition are not well sampled by the repertoire sequencing experiments, making their inference potentially unreliable. Our estimates also rely on the assumption that the dynamics of germinal centers are fast compared to the timescales considered for repertoire evolution. However, germinal centers can last for a few weeks and in some cases for much longer periods [54], and can be seeded by existing memory clones.

Our approach applied to memory-derived plasmablasts infers that, over the course of a year, 1% of memory cells differentiate into plasmablasts. This relatively high rate suggests a scenario in which differentiation is not just the result of isolated events triggered by antigen restimulation but is instead continual and widespread [24, 27, 28]. However, the process is clone- and thus probably antigen-dependent, as evidenced by the large fluctuations that we inferred (small θ_p). The model, however, ignores possible variability and heterogeneity in the differentiation rate across clones and time, as well as heterogeneity in clone degradation of both memory cells and plasmablasts, and folds these variations into the parameters θ_m and θ_p. More realistically, the fate of clones and their ability to proliferate and differentiate into plasmablasts depends on their initial affinity to the antigen, as well as the full history of their affinity maturation. Large, long-lived clones may be driven by a variety of slightly different antigens, e.g. variants of the same pathogen that evolve over years or even decades, like respiratory diseases such as influenza or SARS-CoV-2. More refined, antigen-aware analyses would require reliable labeling of antibody lineages by antigen specificity, which is currently impractical, both experimentally and computationally. The model does not account for the possible coupling between the rates of proliferation and differentiation that one could expect from antigen-dependent stimulation. However, such an effect would likely be ob-

scured by correlations between proliferation and differentiation that are already implied by our model over long timescales: clones that have proliferated a lot tend to be bigger and to have produced more plasmablasts, even at a constant differentiation rate. Additional coupling between proliferation and differentiation arguably occurs at smaller timescales. Future models could include and infer such effects using data with shorter sampling intervals.

Phad et al [24] report the stability of memory B cell clonal families over several years. The argument is based on the observation that the clonal overlap between two samples from the same person taken at different times is not much reduced compared to that between two samples taken at the same time. Mikelov et al [25] also report the stability of memory B cell clones on the year timescale of their experiments based on repertoire overlap. These arguments are indirect and limited to the timescale of the experiment. Here we show that this stability extends for much longer than several years, identifying a characteristic decay time of 15 years. This implies long lifespans of several decades or more for sampled clones, much longer than the span of the original experiments. We also show that this timescale is the same for both individuals, even though they were sampled 6 and 10 years apart, and is consistent with our analysis of data from Mikelov [25], whose donors were sampled 1 year apart, suggesting universality of this timescale. While previous work [24] reports a significant decrease of the number of shared clones for small compared to large clones in samples collected 6 and 10 years apart, we show that this decrease is largely due to sampling biases. We predict that even sampled singletons likely belong to clones large enough to persist over ~ 20 years. In summary, our analysis allows us to put concrete values on the lifetime of clones and memory-to-plasmablast differentiation, extending the list of precise numbers known in biology and immunology [46, 55], characterizing processes that occur on timescales longer than sampled by experiments. The model also allows us to estimate other biologically relevant quantities that are obtainable only through an inference-based approach, such as repertoire turnover, cell proliferation, and germinal-center-related events. Phad et al [24] estimate that 20% of circulating plasmablasts derive from long term memory B cell families, based on the clonal overlap between the two subsets. Our inference approach revealed that this number (ϕ in our notation), cannot be estimated accurately from data. However, we can put a lower bound of 30%. They speculate that sustained plasmablast production may be antigen independent and come from polyclonaly stimulated memory B cells. Our analysis also shows sustained and continuous plasmablast production, however the large heterogeneity in future clone sizes suggests external differentiating factors.

A possible extension of our framework can include subpopulations and could be interesting in the context of memory-cell stability. For example, in [25], from which *Dataset B* is derived, two clusters of low and high temporal persistence were identified. Our analysis ignored these two subgroups, characterizing only the average behavior of the repertoire. Generalizing our approach to multiple subpopulations would increase the number of parameters and the complexity of the model, making the inference harder and leaving some parameters potentially unidentifiable. One could also determine subgroups based on objective observables, such as the isotype, gene expression signatures, sequence features of the B-cell receptor, or phylogenetic properties of the clonal family, run the inference on each subgroup, and compare the learned parameters across them. The method could also be applied to antigen-specific subsets, such as measles, tetanus and influenza clones reported in Phad et al [24], given that these subsets are large enough, and used to predict the evolution of the corresponding antibody titers.

The models inferred using our method reproduce very accurately the statistics of longitudinal repertoire data in healthy conditions. This suggests the possibility of using them as null models to detect and describe pathological or unexpected variations. This approach could be integrated into methods to study clonal expansion and subsequent contraction following immune stimuli, including vaccination [28, 56, 57], or to characterize the unique dynamics of clones interacting with chronic diseases such as HIV [21, 23].

IV. METHODS

A. Inference of clonal families

To group sequences originating from the same naive ancestor into clones, we mainly used the HILARy software [35]. Its computation uses both the heavy and light chains when the pairing is accessible (*Dataset A*), while only the heavy chains for *Dataset B*. Compared to standard software, HILARy uses information on mutations in the V-gene region and optimizes the clustering threshold for each class defined by V,J gene usage and a CDR3 length. Since we need to track longitudinally the same clone at different time points and for different cell types (memory or plasmablast), we performed clone clustering by merging all the samples of the same donor.

To test the robustness of the inference with respect to changes of the clonal assignment algorithm, we also repeated part of the analysis using the software provided by the Change-O toolkit [58], using a threshold of 0.15 on the length-normalized Hamming distance.

B. Inferring a dynamical model from noisy experimental samples

We denote the list of true clonal abundances at a given time point t_k by the vector $\mathbf{n}'^k = \{n'^k_1, n'^k_2, \ldots\}$. The corresponding frequencies read $f'_i = n'_i/N_{\text{cell}}$, where $N_{\text{cell}} = \sum_i n'_i$ is the total number of cells in the body.

The temporal evolution of these abundances is specified by a stochastic dynamical model that depends on some parameters $\mathbf{p}_{\rm dyn}$. The model gives, for the clone i, the probability of observing its frequencies at t_1 and t_2: $P_{\rm dyn}(f'^1_i, f'^2_i|\mathbf{p}_{\rm dyn})$. Since we are interested in studying long term dynamics and not a response to acute stimulations, we assume that the model is in a stationary state and that clones are independent.

Our knowledge about count variations comes from longitudinal sequencing experiments, which provide a very small fraction of the immune repertoire. We define a noise model that provides the experimental count n_i given a real frequency: $P_{\rm noise}(n_i|f'_i, \mathbf{p}_{\rm noise})$. In these notations and throughout, the prime indicates true values for the entire organism, while experimental observations are labeled without prime.

Based on the approach of previous work [29–31], the inference scheme combines the dynamical model and the noise model to write down the probability of observing experimental counts at different time points by integrating over the true frequencies f':

$$P(n_i^1, n_i^2|\mathbf{p}_{\rm noise}, \mathbf{p}_{\rm dyn}) = \int df'^1 df'^2 P_{\rm dyn}(f'^1, f'^2|\mathbf{p}_{\rm dyn}) \times P_{\rm noise}(n_i^1|f'^1, \mathbf{p}_{\rm noise}) P_{\rm noise}(n_i^2|f'^2, \mathbf{p}_{\rm noise}), \tag{1}$$

Using this probability, under the assumption of independent clones, one can then write the log-likelihood of observing the dataset at the different time points:

$$\mathcal{L}(\mathbf{p}_{\rm noise}, \mathbf{p}_{\rm dyn}) = \sum_i \log P(n_i^1, n_i^2|n_i^1+n_i^2 > 0, \mathbf{p}_{\rm noise}, \mathbf{p}_{\rm dyn}), \tag{2}$$

where the probabilities are conditioned on clone i to be present in the dataset. The maximization of the likelihood allows us to find the best set of parameters that describe the data. Practically, we first learn the noise parameters $\mathbf{p}_{\rm noise}$ by using experimental replicates as explained in the next section, and, secondly, we learn the parameters of the dynamics $\mathbf{p}_{\rm dyn}$ by plugging the optimal $\mathbf{p}^*_{\rm noise}$ into Eq. 1.

C. Noise models

Because repertoire sequencing experiments are based on samples representing a small part of the entire organism, and because of experimental noise, experimental counts are a noisy function of the true clonal frequencies. We describe this noisy relationship through three alternative models that we present in order of complexity, i.e., the number of adjustable parameters they have. In a pure sampling regime, the experimental-count probability of the clone i is given by a Poisson distribution with average $\mu_i = Mf'_i$, where $M = \sum_i n_i$ is the total number of counts in the experiment (single cell, reads, or unique molecular identifier, depending on the case). This parameter-free model describes a scenario in which there is no other noise than from sampling. The other two models are based on the negative binomial distribution, commonly used in the presence of amplification biases in RNA sequencing data [59] and repertoire sequencing [29, 30]. We parametrize this distribution by fixing the mean to $\mu_i = Mf'_i$ as before and writing down the variance σ_i^2 as an increasing function of the mean, with two free parameters a and b:

$$\sigma_i^2 = \mu_i + a(\mu_i)^b. \tag{3}$$

When $a = 0$ we get back the Poisson distribution. We consider two options: one in which the exponent b is fixed to 1, i.e., the variance is proportional to the average with a factor of $1 + a$, and the full model with the two parameters (a, b) to learn.

This noise model is inferred by using experimental replicates performed at the same time point. Given the experimental count n_i^r for the clone i in the replicate r, we can write down the probability of observing the clone in two different replicates:

$$P(n_i^1, n_i^2|\mathbf{p}_{\rm stat}, \mathbf{p}_{\rm noise}) = \int df' \; P_{\rm stat}(f'|\mathbf{p}_{\rm stat}) \times P_{\rm noise}(n_i^1|\mu = f'M, \mathbf{p}_{\rm noise}) P_{\rm noise}(n_i^2|\mu = f'M, \mathbf{p}_{\rm noise}), \tag{4}$$

where $P_{\rm stat}(n|\mathbf{p}_{\rm stat})$ is the stationary distribution for the repertoire abundances, which we assume to be power-law distributed, $P_{\rm stat}(f'|\mathbf{p}_{\rm stat}) \propto 1/f'^{\alpha+1}$, in agreement with experimental observations [33, 60] and previous analyses [31, 32, 61]. This distribution depends on two free parameters: the power law exponent α and a minimal frequency $f_{\rm min}$ to ensure integrability. By using the Eq. 4, one can obtain the likelihood of observing a set of replicates using Eq. 1 and Eq. 2 but with $t_1 = t_2$. Sec S2 provides specific details for its computation and maximization.

$N_{\rm cell}$	Number of memory B-cells in the body
M	Number of cells in the experiment
$N_{\rm clone}$	Number of memory clones in the body
α	Power law exponent of the cumulative clone count distributio
$f_{\rm min}$	Minimal clonal frequency $= 1/N_{\rm cell}$
a, b	Negative Binomial noise parameters (Eq. 3)
τ_m	Decay time of the GBM for memory clones
θ_m	Fluctuation time of the GBM for memory clones
s	Rate of introduction of new memory clones
n_0	Initial number of cells of a new clone
ρ	Differentiation rate from memory to plasmablast
τ_p	Decay time of the GBM for plasmablasts
θ_p	Fluctuation time of the GBM for plasmablasts
ϕ	Fraction of plasmablasts with memory origin
r	Plasmablast to memory cell ratio in the body

TABLE III: List of parameters.

D. Geometric Brownian Motion for memory cell dynamics

The minimal model to describe the dynamics of memory B cells in healthy condition is the Geometric Brownian Motion (GBM), according to which abundances of cells decrease exponentially on average and are subject to geometric fluctuations:

$$\frac{dn'_i}{dt} = -\frac{n'_i}{\tau_m} + \frac{n'_i}{\sqrt{\theta_m}}\eta_i(t), \tag{5}$$

where τ_m is the timescale for the average exponential degradation, θ_m quantifies the amplitude of fluctuations, and $\eta_i(t)$ is a normalized Gaussian white noise implemented with the Itô convention. Clones go extinct when they hit the absorbing boundary $n' = 1$, and new clones are introduced with a constant rate, s, starting with an initial abundance n_0. The balance of extinction and new introduction leads to a stationary condition, with an abundance distribution that has a power law tail $p(n') \propto 1/n'^{\alpha+1}$ (see [32] and Sec. S2E), where the exponent is connected with the model parameters through $\alpha = 2\theta_m/\tau_m$. The power law behavior and its simplicity are the main reasons for choosing this model. In addition, it has been shown that when observing two time points sufficiently far apart in time, i.e. longer than a few weeks, more refined models, including antigen recognition and clonal expansion, can be well approximated by the GBM [32].

Eq. 5 defines a dynamical model P_{dyn} (see Sec. S2E). Combining it with the noise model P_{noise} allows us to run the inference machinery described above (Eqs. 1, 2) and to infer the biologically informative parameters of the GBM, as discussed in Sec. II D.

E. Coupling plasmablast and memory cell dynamics

Memory cells can differentiate into plasmablasts, providing a rapid response against pathogens encountered in previous infections. We denote by ρ this differentiation rate, and we let the dynamics of the number of plasmablasts in the clone i, k'_i, follow a Geometric Brownian Motion as for memory cells:

$$\frac{dk'_i}{dt} = \rho n'_i - \frac{k'_i}{\tau_p} + \frac{k'_i}{\sqrt{\theta_p}}\xi_i(t). \tag{6}$$

Plasmablast clones degrade exponentially on average with a time τ_p and are subject to fold fluctuations over time scale θ_p. By coupling Eqs. 5 and 6, one obtains a dynamical model and can learn the plasmablast parameters given the memory parameters learned previously. It turns out that it is not possible to learn the three new parameters simultaneously. This can be seen by writing the equation for the true frequencies $g'_i = k'_i/K'_m$, where K'_m is the total number of plasmablasts generated by memory cells, and observing that it loses the dependency on ρ (see Sec. S4). This indeterminacy can be solved by finding the relation between the total number of cells at stationary state, $K'_m = \rho\tau_p N_{\text{cell}}$, and comparing it with the empirical ratio of these two numbers. In particular, we call $\phi = K'_m/K'$ the fraction of plasmablasts that differentiated from memory cells compared to the total population of plasmablasts K' (some plasmablasts can differentiate from naive cells), and $r = K'/N_{\text{cell}}$ the plasmablast-to-memory ratio. This leads to $K'_m = r\phi N_{\text{cell}}$, and $\rho = r\phi/\tau_p$, where τ_p is one of the parameters to fit in the optimization of the likelihood, and r and ϕ are directly given by empirical estimates from the repertoires.

F. Code Availability

The code used for processing the immune-sequencing repertoire datasets, for performing the inference, and for generating the figures is available at the repository `https://github.com/statbiophys/memory_plasmab_dynamics`

Acknowledgements

This work was partially supported by the ANR-19-CE45-0018 “RESP-REP" and ANR-22-CE95-0005-01 “DISTANT” grants from the Agence Nationale de la Recherche.

[1] S Tonegawa, Somatic generation of antibody diversity. *Nature* **302**, 575–581 (1983).

[2] FW Alt, et al., VDJ recombination. *Immunology today* **13**, 306–314 (1992).

[3] G Lythe, RE Callard, RL Hoare, C Molina-París, How many TCR clonotypes does a body maintain? *Journal of theoretical biology* **389**, 214–224 (2016).

[4] G Altan-Bonnet, T Mora, AM Walczak, Quantitative immunology for physicists. *Physics Reports* **849**, 1–83 (2020).

[5] IC MacLennan, Germinal centers. *Annual review of immunology* **12**, 117–139 (1994).

[6] GD Victora, MC Nussenzweig, Germinal centers. *Annual review of immunology* **40**, 413–442 (2022).

[7] Y Elhanati, et al., Inferring processes underlying B-cell repertoire diversity. *Philosophical Transactions of the Royal Society B: Biological Sciences* **370**, 20140243 (2015).

[8] KB Hoehn, A Fowler, G Lunter, OG Pybus, The diversity and molecular evolution of B-cell receptors during

infection. *Molecular biology and evolution* **33**, 1147–1157 (2016).

[9] MJ Shlomchik, F Weisel, Germinal center selection and the development of memory B and plasma cells. *Immunological reviews* **247**, 52–63 (2012).

[10] N Lam, Y Lee, DL Farber, A guide to adaptive immune memory. *Nature Reviews Immunology* **24**, 810–829 (2024).

[11] SD Boyd, et al., Measurement and clinical monitoring of human lymphocyte clonality by massively parallel VDJ pyrosequencing. *Science translational medicine* **1**, 12ra23–12ra23 (2009).

[12] A Six, et al., The past, present, and future of immune repertoire biology–the rise of next-generation repertoire analysis. *Frontiers in immunology* **4**, 413 (2013).

[13] WS DeWitt, et al., A public database of memory and naive B-cell receptor sequences. *PloS one* **11**, e0160853 (2016).

[14] C Soto, et al., High frequency of shared clonotypes in human B cell receptor repertoires. *Nature* **566**, 398–402 (2019).

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

[16] CO McCoy, et al., Quantifying evolutionary constraints on B-cell affinity maturation. *Philosophical Transactions of the Royal Society B: Biological Sciences* **370**, 20140244 (2015).

[17] F Horns, C Vollmers, CL Dekker, SR Quake, Signatures of selection in the human antibody repertoire: Selective sweeps, competing subclones, and neutral drift. *Proceedings of the National Academy of Sciences* **116**, 1261–1266 (2019).

[18] BJ Olson, FA Matsen IV, The Bayesian optimist's guide to adaptive immune receptor repertoire analysis. *Immunological reviews* **284**, 148–166 (2018).

[19] DK Ralph, FA Matsen IV, Using B cell receptor lineage structures to predict affinity. *PLOS Computational Biology* **16**, e1008391 (2020).

[20] N Spisak, AM Walczak, T Mora, Learning the heterogeneous hypermutation landscape of immunoglobulins from high-throughput repertoire data. *Nucleic acids research* **48**, 10702–10712 (2020).

[21] A Nourmohammad, J Otwinowski, M Łuksza, T Mora, AM Walczak, Fierce selection and interference in B-cell repertoire response to chronic HIV-1. *Molecular biology and evolution* **36**, 2184–2194 (2019).

[22] KB Hoehn, et al., Human B cell lineages engaged by germinal centers following influenza vaccination are measurably evolving. *bioRxiv* (2021).

[23] A Mazzolini, T Mora, AM Walczak, Inspecting the interaction between HIV and the immune system through genetic turnover. *Philos. Trans. R. Soc. B* (2022).

[24] GE Phad, et al., Clonal structure, stability and dynamics of human memory B cells and circulating plasmablasts. *Nature immunology* **23**, 1076–1085 (2022).

[25] A Mikelov, et al., Memory persistence and differentiation into antibody-secreting cells accompanied by positive selection in longitudinal BCR repertoires. *Elife* **11**, e79254 (2022).

[26] AF Ochsenbein, et al., Protective long-term antibody memory by antigen-driven and T help-dependent differentiation of long-lived memory B cells to short-lived plasma cells independent of secondary lymphoid organs. *Proceedings of the National Academy of Sciences* **97**, 13263–13268 (2000).

[27] NL Bernasconi, E Traggiai, A Lanzavecchia, Maintenance of serological memory by polyclonal activation of human memory B cells. *Science* **298**, 2199–2202 (2002).

[28] F Horns, CL Dekker, SR Quake, Memory B cell activation, broad anti-influenza antibodies, and bystander activation revealed by single-cell transcriptomics. *Cell reports* **30**, 905–913 (2020).

[29] M Puelma Touzel, AM Walczak, T Mora, Inferring the immune response from repertoire sequencing. *PLOS Computational Biology* **16**, e1007873 (2020).

[30] MB Koraichi, MP Touzel, A Mazzolini, T Mora, AM Walczak, NoisET: Noise Learning and Expansion Detection of T-Cell Receptors. *The Journal of Physical Chemistry A* **126**, 7407–7414 (2022).

[31] M Bensouda Koraichi, S Ferri, AM Walczak, T Mora, Inferring the T cell repertoire dynamics of healthy individuals. *Proceedings of the National Academy of Sciences* **120**, e2207516120 (2023).

[32] J Desponds, T Mora, AM Walczak, Fluctuating fitness shapes the clone-size distribution of immune repertoires. *Proceedings of the National Academy of Sciences* **113**, 274–279 (2016).

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

[34] T Mora, AM Walczak, Quantifying lymphocyte receptor diversity in *Systems Immunology.* (CRC Press), pp. 183–198 (2018).

[35] N Spisak, G Athènes, T Dupic, T Mora, AM Walczak, Combining mutation and recombination statistics to infer clonal families in antibody repertoires. *eLife* **13**, e86181 (2024).

[36] Y Pan, et al., The Poisson distribution model fits UMI-based single-cell RNA-sequencing data. *BMC bioinformatics* **24**, 256 (2023).

[37] S Lazzardi, et al., Emergent statistical laws in single-cell transcriptomic data. *Physical Review E* **107**, 044403 (2023).

[38] M Seifert, R Küppers, Human memory B cells. *Leukemia* **30**, 2283–2292 (2016).

[39] R Sender, et al., The total mass, number, and distribution of immune cells in the human body. *Proceedings of the National Academy of Sciences* **120**, e2308511120 (2023).

[40] HC Matz, KM McIntire, AH Ellebedy, Persistent germinal center responses: slow-growing trees bear the best fruits. *Current opinion in immunology* **83**, 102332 (2023).

[41] H Tabibian-Keissar, et al., Aging affects B-cell antigen receptor repertoire diversity in primary and secondary lymphoid tissues. *European journal of immunology* **46**, 480–492 (2016).

[42] R Carsetti, et al., Comprehensive phenotyping of human peripheral blood B lymphocytes in healthy conditions. *Cytometry Part A* **101**, 131–139 (2022).

[43] L Khodadadi, Q Cheng, A Radbruch, F Hiepe, The maintenance of memory plasma cells. *Frontiers in immunology* **10**, 721 (2019).

[44] SL Nutt, PD Hodgkin, DM Tarlinton, LM Corcoran, The generation of antibody-secreting plasma cells. *Nature Reviews Immunology* **15**, 160–171 (2015).

[45] RJ De Boer, H Mohri, DD Ho, AS Perelson, Turnover

Rates of B Cells, T Cells, and NK Cells in Simian Immunodeficiency Virus-Infected and Uninfected Rhesus Macaques. *The Journal of Immunology* **170**, 2479–2487 (2003).

[46] J Cosgrove, LS Hustin, RJ de Boer, L Perié, Hematopoiesis in numbers. *Trends in Immunology* **42**, 1100–1112 (2021).

[47] A Krueger, N Ziętara, M Łyszkiewicz, T cell development by the numbers. *Trends in immunology* **38**, 128–139 (2017).

[48] R Sender, R Milo, The distribution of cellular turnover in the human body. *Nature medicine* **27**, 45–48 (2021).

[49] IJ Amanna, NE Carlson, MK Slifka, Duration of humoral immunity to common viral and vaccine antigens. *New England Journal of Medicine* **357**, 1903–1915 (2007).

[50] S Crotty, et al., Cutting edge: long-term B cell memory in humans after smallpox vaccination. *The Journal of Immunology* **171**, 4969–4973 (2003).

[51] L Westera, et al., Lymphocyte maintenance during healthy aging requires no substantial alterations in cellular turnover. *Aging cell* **14**, 219–227 (2015).

[52] MP Cancro, MM Tomayko, Memory B cells and plasma cells: The differentiative continuum of humoral immunity. *Immunological reviews* **303**, 72–82 (2021).

[53] M Molari, K Eyer, J Baudry, S Cocco, R Monasson, Quantitative modeling of the effect of antigen dosage on B-cell affinity distributions in maturating germinal centers. *Elife* **9**, e55678 (2020).

[54] L Mesin, J Ersching, GD Victora, Germinal center B cell dynamics. *Immunity* **45**, 471–482 (2016).

[55] R Milo, P Jorgensen, U Moran, G Weber, M Springer, Bionumbers—the database of key numbers in molecular and cell biology. *Nucleic acids research* **38**, D750–D753 (2010).

[56] JD Galson, et al., B-cell repertoire dynamics after sequential hepatitis B vaccination and evidence for cross-reactive B-cell activation. *Genome medicine* **8**, 68 (2016).

[57] KB Hoehn, et al., Cutting edge: distinct B cell repertoires characterize patients with mild and severe COVID-19. *The Journal of Immunology* **206**, 2785–2790 (2021).

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

[59] MI Love, W Huber, S Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* **15**, 1–21 (2014).

[60] N Jiang, et al., Lineage structure of the human antibody repertoire in response to influenza vaccination. *Science translational medicine* **5**, 171ra19–171ra19 (2013).

[61] V Chardès, M Vergassola, AM Walczak, T Mora, Affinity maturation for an optimal balance between long-term immune coverage and short-term resource constraints. *Proceedings of the National Academy of Sciences* **119**, e2113512119 (2022).

[62] GX Zheng, et al., Massively parallel digital transcriptional profiling of single cells. *Nature communications* **8**, 14049 (2017).

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

[64] MP Lefranc, et al., IMGT®, the international ImMuneGeneTics information system®. *Nucleic acids research* **37**, D1006–D1012 (2009).

[65] DA Bolotin, et al., MiXCR: software for comprehensive adaptive immunity profiling. *Nature methods* **12**, 380–381 (2015).

[66] AN Borodin, P Salminen, *Handbook of Brownian motion-facts and formulae.* (Birkhäuser), (2012).

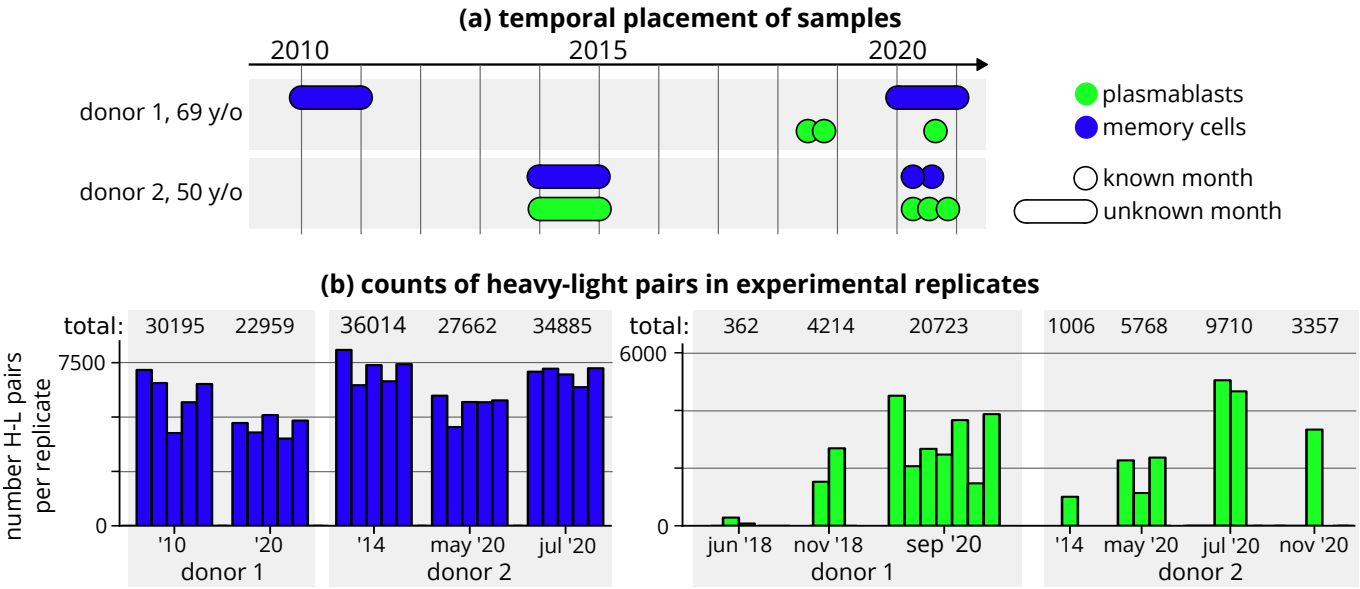
SUPPLEMENTARY MATERIAL

S1. DATASETS

A. Single-cell *Dataset A*

The single-cell immune sequencing *Dataset A* is described in [24]. B cells were obtained with the high-throughput single-cell 10X Genomics platform and available under ArrayExpress accession *E-MTAB-11174* and *E-MTAB-11697*. This dataset consists of longitudinal sequences of two male donors (69 and 50 years old at the last data point) that span a time window of several years and that are divided into memory cells and plasmablasts, Fig. S1a. Memory cells and plasmablasts were identified and sorted based on surface markers and their identity was verified using transcriptome lineage markers. Each sample is divided into replicates (10X lanes) that are sequenced in parallel. The number of replicates per sample is 5 for memory cells and variable, from 1 to 7, for plasmablasts.

We assembled the `fastq` raw files using Cell Ranger (command vdj) [62], and then the V(D)J sequences are aligned and annotated using the Change-O toolkit [58], which employs IgBLAST [63] for the alignment to germline sequences provided by the IMGT database [64]. After these steps, we keep only cells that have a clean heavy and light chain pairing. The size of samples divided into replicates are reported in Fig. S1b, with a total of around 30,000 cells per memory sample and a very variable number for plasmablasts (from a few hundred to few tens of thousands). We excluded a sample from the original dataset for donor 1 taken in February 2021 because it contained both memory and plasmablast cells without the possibility of discrimination. Clone clustering is performed with HILARy [35] as described in the Methods section.



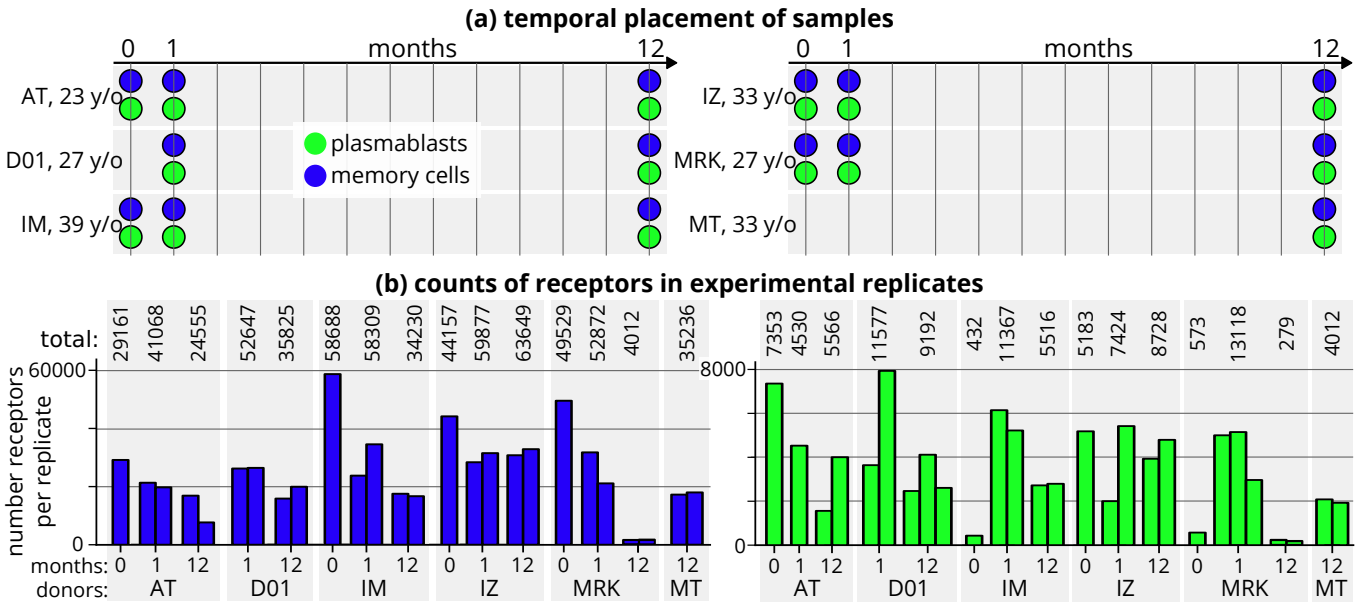
Supplementary Figure S1: ***Dataset A*** [24] (a): dataset samples of the two donors for memory cells (blue) and plasmablasts (green). For some samples the metadata provides the year and month of the experiment (circles), while for others only the year (elongated points). (b): counts of cells that contain a heavy-light sequence for each experimental replicates of each sample. On the left memory cell samples (blue), on the right plasmablasts (green).

B. cDNA amplification *Dataset B*

Dataset B [25] is from an IGH repertoire sequencing experiment with amplification of cDNA molecules (5'-RACE based protocol) using Unique Molecular Identifiers. The dataset is available in the ArrayExpress repository under identifier *E-MTAB-11193*. Using fluorescence-activated cell sorting (FACS), cells were sorted into memory cells and plasmablasts for six donors, aged 23 to 39 years old (Fig. S2a). Samples are longitudinal for a maximum of 3 time

points separated by 1 and 11 months, and can have up to three experimental replicates, see Fig. S2a. For the donor MT, only the samples at the last time point are available.

We performed the raw data processing by using MiXCR [65] using the preset `mikelov-et-al-2021`. For practical reasons, the obtained corrected sequences were also aligned with IgBLAST [63] to germline sequences provided by the IMGT database [64]. The sizes of samples divided into replicates are reported in Fig. S2b, with a variable number of receptors: an average of 25,000 for memory cells and an average of 3500 for plasmablasts. As for the previous dataset, the clonal families were built with HILARy [35], with the difference that here we used the information only on the heavy chain.



Supplementary Figure S2: ***Dataset B*** [25] (a): dataset samples of the six donors for memory cells (blue) and plasmablasts (green). (b): counts of receptors for each experimental replicates for each sample. On the left memory cell samples (blue), on the right plasmablasts (green).

S2. LIKELIHOOD COMPUTATION AND MAXIMIZATION FOR THE NOISE MODEL

A. Inference of the noise model with replicates

Inference of the noise model is performed in each sample by exploiting its division into replicates. We assume that the full repertoire of the donor is composed of N_{clone} different unique clones, each of them with a given count n'_i, $i = 1, \dots, N_{\text{clone}}$. By defining the total number of counts as $N_{\text{cell}} = \sum_{i=1}^{N_{\text{clone}}} n'_i$, we can define the corresponding frequencies as $f'_i = n'_i / N_{\text{cell}}$. As mentioned in the main text, we assume that the stationary distribution of these counts/frequencies is power-law distributed:

$$f'_i \sim P_{\text{stat}}(f' | \mathbf{p}_{\text{stat}}) = \alpha (f_{\min})^{\alpha} f'^{-\alpha-1}, \quad f_{\min} \leq f' \leq 1, \tag{S1}$$

where we consider $f_{\min} = 1/N_{\text{cell}} \ll 1$ and $\alpha > 1$ (consistent with observations).

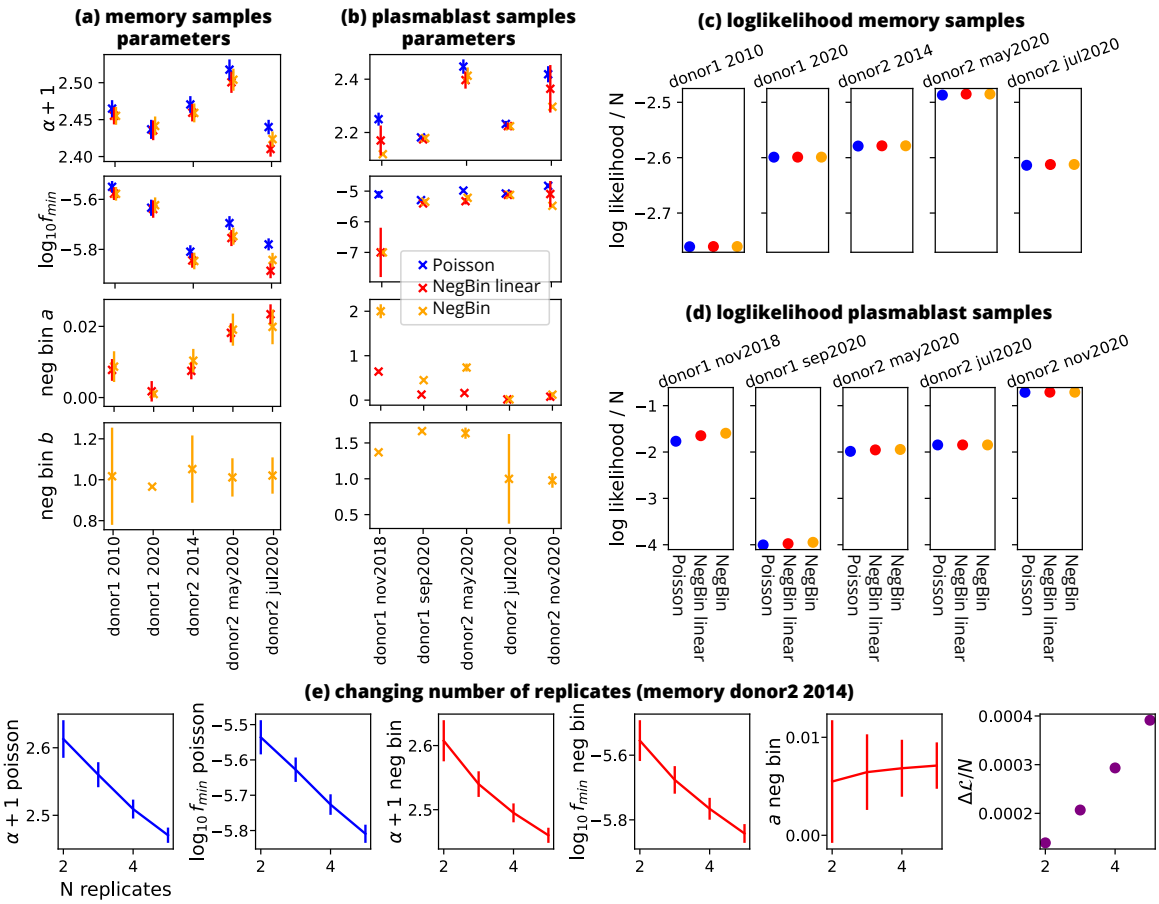
This distribution depends on two parameters, $\mathbf{p}_{\text{stat}} = (\alpha, f_{\min})$. Notice that, overall, the repertoire frequencies are characterized by three parameters: α, $f_{\min}$, and N_{clone}. However, they will be reduced to 2 since the frequency normalization imposes the constraint $\sum_i^{N_{\text{clone}}} f'_i = 1$. For easier computations, we impose the average of this constraint, as also used in a previous work with a similar setting [29]:

$$N_{\text{clone}} \langle f' \rangle = 1, \tag{S2}$$

where the average is over the distribution P_{stat} in Eq. S1.

A sequencing experiment consists of sampling these frequencies to obtain experimental counts of clones. We call $\mathbf{n}^r = \{n_1^r, n_2^r, \dots\}$ the counts of clones in the replicate r, where N_r is the total number of unique clones in the replicate. The total number of cells in the replicate is $M_r = \sum_{i=1}^{N_r} n_i^r$.

We now want to write the probability of sampling n_i^r cells of the clone i in the replicate r, i.e., the noise model $P_{\text{noise}}(n_i^r|\mathbf{p}_{\text{noise}})$. In general, we want this probability to have an unbiased average $\mu_i^r = Mf_i'$. We consider two functional forms for the noise probability. The first is a Poisson distribution, which is completely defined by its mean and does not need any additional parameter. The second is a negative binomial distribution parametrized by choosing the variance as $\sigma^2{}_{NB} = \mu + a\mu^b$. We test a first case in which we fix $b = 1$, i.e., there is one free parameter $\mathbf{p}_{\text{noise}} = a$, and variance is proportional to the average with a factor $1 + a$. For the second case we let the inference learn both the parameters $\mathbf{p}_{\text{noise}} = (a, b)$.

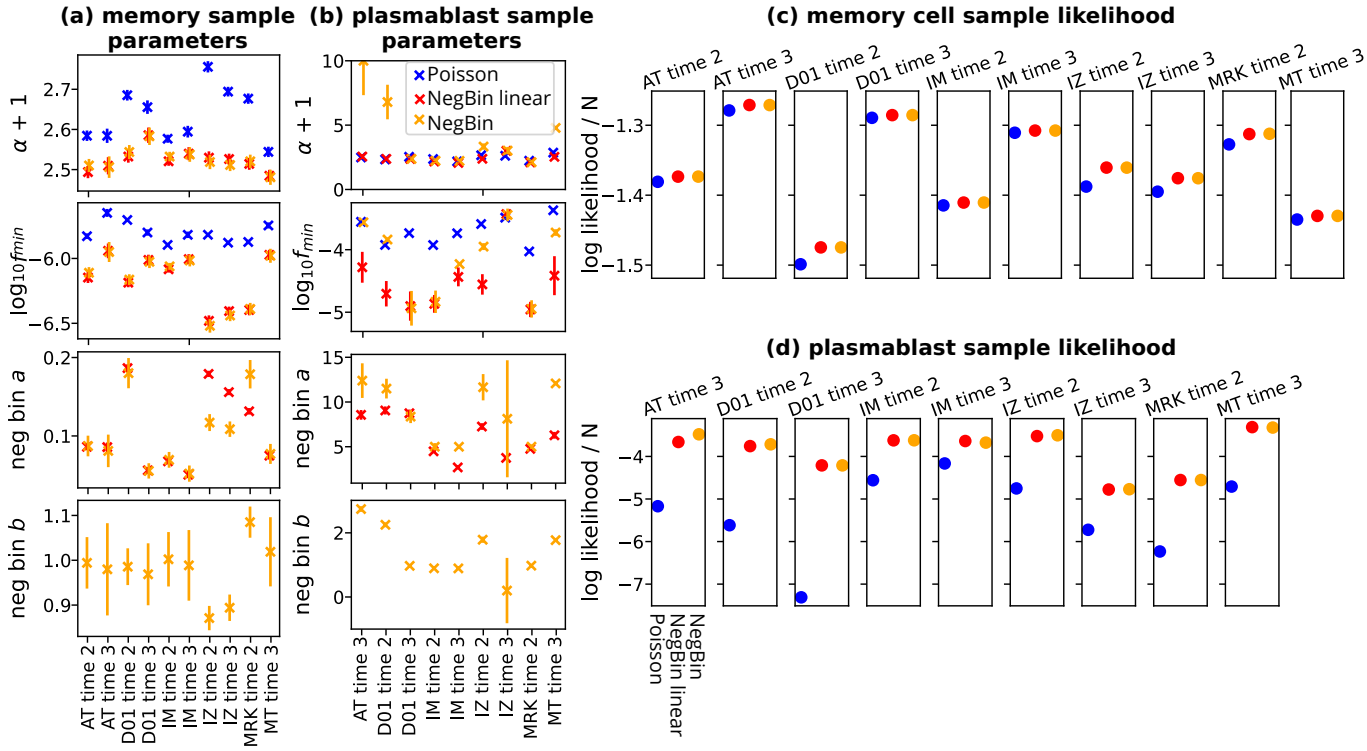


Supplementary Figure S3: **Results of the noise inference for *Dataset A*.** (a), (b): parameters learned in the three inference models represented with three different colors for samples of memory B-cells (a) and plasmablasts (b). The error bars are computed with the Hessian of the log-likelihood. (c), (d): values of the log-likelihood normalized by the number of clones for the three models (on the x-axis), for each sample, for memory B-cells (c) and plasmablasts (d). We run the noise inference of the Poisson and the linear Negative Binomial models by varying the number of replicates of donor 2, 2014 sample, *Dataset A*. (e): dependence on the number of replicates of the inference (donor 2 2014, memory cells). For a given number of replicates $(2, 3, 4, 5)$ we run all the combinations among the 5 total replicates that are present. We report the average learned parameter and its average error across all the combinations. The two leftmost plots (in blue) show the Poisson parameters, the next three plots (in red) show the linear Negative Binomial parameters, and the rightmost plot is the difference of the log-likelihood per sample between the Negative Binomial and the Poisson model.

B. Likelihood maximization for the noise model

Given a dataset with R replicates, $\{\mathbf{n}^1, \ldots, \mathbf{n}^R\}$, we start by computing the probability that the cells of the clone i have been observed with abundance $\{n_i^1, \ldots, n_i^R\}$ in the replicates by using the power law distribution (S1) and one of the alternative noise probabilities discussed before:

$$P(n_i^1, \ldots, n_i^R|\mathbf{p}) = \int_{f_{\min}}^{1} df' \; P_{\text{stat}}(f'|\mathbf{p}_{\text{stat}}) \prod_{r=1}^{R} P_{\text{noise}}(n_i^r|\mu = M_r f', \mathbf{p}_{\text{noise}}), \tag{S3}$$



Supplementary Figure S4: **Results of the noise inference for *Dataset B*.** (a), (b): parameters learned in the three inference models represented with three different colors for samples of memory B-cells (a) and plasmablasts (b). The error bars are computed with the Hessian of the log-likelihood. (c), (d): values of the log-likelihood normalized by the number of clones for the three models (on the x-axis), for each sample, for memory B-cells (c) and plasmablasts (d).

where $\mathbf{p} = \mathbf{p}_{\text{stat}} \cup \mathbf{p}_{\text{noise}}$. That equation generalizes Eq. 4 to more than 2 replicates.

The next step is to compute the log-likelihood of observing the counts of the N clones in the dataset, which is based on the assumption that they are independent:

$$\mathcal{L}(\mathbf{n}^1, \ldots, \mathbf{n}^R|\mathbf{p}) = \sum_{i=1}^{N} \log P\left(n_i^1, \ldots, n_i^R \,\middle|\, \sum_r n_i^r > 0, \mathbf{p}\right) = \sum_{i=1}^{N} \log\left(\frac{P(n_i^1, \ldots, n_i^R|\mathbf{p})}{1 - P(\mathbf{0}|\mathbf{p})}\right). \tag{S4}$$

Notice that we conditioned the fact that a clone is present, i.e., $\sum_r n_i^r > 0$. The second equality explicitly writes down this probability.

To get the best parameters that best describe the data, the final step is to maximize this likelihood over the set of parameters. We have to remember that there is one constraint, Eq. S2, that has to be satisfied during the maximization:

$$\theta^* = \arg\max_{\theta} \mathcal{L}(\mathbf{n}^1, \ldots, \mathbf{n}^R|\mathbf{p}) \quad \text{satisfying} \quad \langle f' \rangle N_{\text{clone}} = 1. \tag{S5}$$

The results of this optimization are shown in Fig. S3 and S4 for the two datasets and for the three different noise models discussed before. The panels a and b of the figure show the learned parameters $\mathbf{p}$: 2 for the stationary distribution, $\mathbf{p}_{\text{stat}} = (\alpha, f_{\text{min}})$, and the set of model-specific parameters $\mathbf{p}_{\text{noise}}$ (none for the Poisson model, a for the negative binomial with $b = 1$, i.e., linear, and (a, b) for the full negative binomial). Panels c and d compare the log-likelihood per sample of the three models, allowing us to compare their performances and choose the most appropriate depending on the dataset.

C. Numerical implementation of the likelihood maximization for the noise model

The likelihood optimization is computed in Python, using wrapped functions in C++ to numerically solve the integrals for the probability of counts, Eq. S3. The code can be found in the repository `https://github.com/statbiophys/memory_plasmab_dynamics`.

The integrals were solved with a standard trapezoid method using a logarithmic binning in the space of frequencies, with typically 10^4 bins. A useful trick to speed up the computation was to exclude from the integration all the frequency domain that was approximately zero for the integrand $g(f)$ of Eq. S3, since for many cases it is a peaked narrow function. The threshold ϵ_0 to determine if the integrand is approximately zero is chosen as the integrand maximum divided by 10^6. To move from $[f_{\min}, 1]$ to a new integration domain $[f_{\text{left}}, f_{\text{right}}]$ where the integrand is larger than the threshold, we scan the frequency axis by using a bisection algorithm that finds the values for which the integrand crosses the threshold ϵ_0.

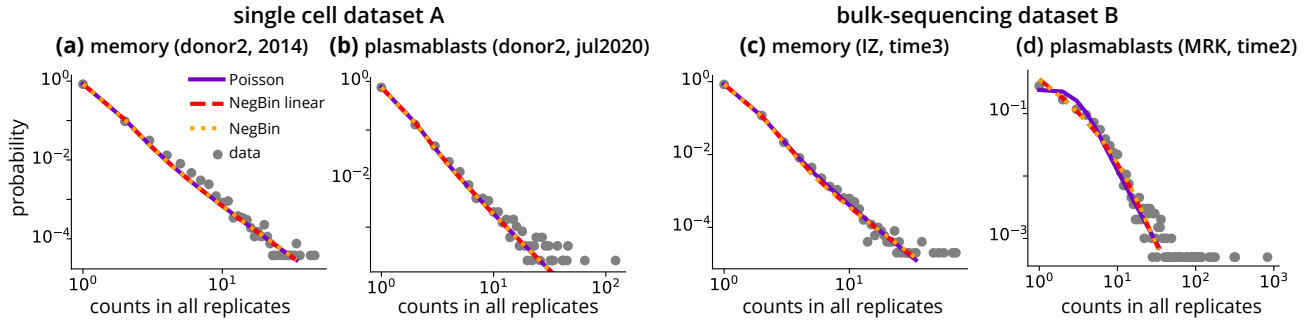
Optimization was performed in python using Brent's method implemented in the *scipy* library, which allows for constrained optimization. To speed up the computation, the summation terms of the likelihood in Eq. S4 are evaluated in parallel across multiple machine cores, using the *multiprocessing* library of python.

D. Errors estimates

An estimate of the uncertainty of the optimal parameters can be obtained by the inverse of the Hessian of the log-likelihood, whose element corresponding to the parameters p_i and p_j is $H_{i,j} = -\partial_{p_i}\partial_{p_j}\mathcal{L}(\mathbf{n}^1, \ldots, \mathbf{n}^R|\mathbf{p})$. In a likelihood maximization without constraints, the error associated with the parameter p_i is $\sigma_i = \sqrt{H^{-1}_{i,i}}$. However, we have to satisfy the constraint $C(\mathbf{p}) = \langle f \rangle N - 1 = 0$. To correctly include this in the uncertainty estimate, we need to project the Hessian on the hyperplane tangent to the constraint at the solution p^*. This can be done by using the projection matrix $P = I - \mathbf{u}^T\mathbf{u}$, where I is the identity matrix and $\mathbf{u} = \nabla C(\mathbf{p}^*)/|\nabla C(\mathbf{p}^*)|$ is the normal unit vector to the constraint. The local projection of the Hessian matrix along the constraint then reads $\hat{H} = PHP$. The projected Hessian is no longer invertible since one eigenvalue is zero along the direction perpendicular to the constraint. However, the error can be computed by summing all the contributions along each eigenvector of $\hat{H}$ and excluding that direction. If we call λ_k the eigenvalues of $\hat{H}$, putting the zero at $k = 0$, i.e., $\lambda_0 = 0$, and Q the matrix having the eigenvectors of $\hat{H}$ as columns, the error associated with the parameter p_i is

$$\sigma_i = \sqrt{\sum_{k\neq 0} \frac{Q_{i,k}Q^{-1}_{k,i}}{\lambda_k}}.$$

On a technical level, we realized that a numerical computation that evaluates $\mathcal{L}$ in $\mathbf{p}^*$, $\mathbf{p}^* - \delta\mathbf{p}$, $\mathbf{p}^* + \delta\mathbf{p}$ and uses the discrete formula for the second derivatives was leading to very unstable results, extremely sensitive to the size of $\delta\mathbf{p}$. We decided then to compute analytically the first and second derivatives of Eq. S3 with respect to all the parameters and evaluate the Hessian using those exact expressions.



Supplementary Figure S5: **Clone size distributions predicted by the noise model.** The empirical clone size distribution is compared with the prediction of the three inferred noise models: Poisson model, negative binomial with a fixed parameter $b = 1$ and full negative binomial. We show the distribution of total counts in a sample n, which is the sum of the clone counts in each replicate: $n = \sum_r n^r$. (a): memory repertoire of donor 2 of *Dataset A*, 2014 timepoint, (b): plasmablast repertoire of donor 2 of *Dataset A*, July 2020 timepoint, (c): memory repertoire of donor IZ of *Dataset B*, third time point), (d): plasmablast repertoire of donor MRK of *Dataset B*, second time point).

S3. LIKELIHOOD COMPUTATION AND MAXIMIZATION FOR THE GEOMETRIC BROWNIAN MOTION

A. Geometric Brownian Motion

The stochastic differential equation of the Geometric Brownian Motion (GBM) reads:

$$\frac{dn_i'(t)}{dt} = -\frac{n_i'}{\tau} + \frac{n_i'}{\sqrt{\theta}}\eta_i(t), \tag{S6}$$

where τ is the time scale for the average exponential degradation, θ quantifies the amplitude of fluctuations, and $\eta_i(t)$ is a Gaussian white noise with the Itô convention. Clones go extinct when their abundance goes below 1, which is mathematically modeled by introducing an absorbing boundary at $n' = 1$, and new clones are introduced with a constant rate, s, starting with an initial abundance n_0. This model is fully solvable by taking advantage of the substitution $x = \log n'$, which recovers a classical Brownian Motion for the log-size x.

The Fokker-Plank equation of the log-size reads:

$$\partial_t P(x,t) = \frac{1}{\tau}\partial_x P(x,t) + \frac{1}{2\theta}\partial_x^2 P(x,t) + s\delta(x - x_0),$$

where $x_0 = \log n_0$. This equation admits a stationary distribution that can be computed analytically:

$$P_{\text{stat}}(x) = \tau s e^{-\alpha x} \begin{cases} e^{\alpha x} - 1 & \text{if } x \leq x_0 \\ e^{\alpha x_0} - 1 & \text{if } x > x_0 \end{cases}, \tag{S7}$$

where we introduce $\alpha = 2\theta/\tau$. This function shows a power law tail with exponent $-(\alpha + 1)$ in the limit of large abundance: $P_{\text{stat}}(x) \sim e^{-\alpha x}$ and $P_{\text{stat}}(n') = P_{\text{stat}}(x)/n' \sim n'^{-\alpha-1}$. We should notice that Eq. S7 normalizes to the number of different clones in the repertoire $\int P_{\text{stat}}(x)dx = N_{\text{clone}}$, which allows us to find:

$$N_{\text{clone}} = \tau s \log n_0 = \frac{\alpha - 1}{\alpha}\frac{N_{\text{cell}}}{n_0}\log n_0, \tag{S8}$$

where the second expression is rewritten by using Eq. S9 (see below) to remove the dependency on τ, which is practically useful for estimating this number in the second dataset. We can also find the total number of cells by solving the integral $N_{\text{cell}} = \int n' P_{\text{stat}}(n')dn'$:

$$N_{\text{cell}} = s(n_0 - 1)\left(\frac{1}{\tau} - \frac{1}{2\theta}\right)^{-1}. \tag{S9}$$

Below we derive formulas to predict the survival properties of trajectories in a GBM used in the main text (Fig. 4). The first quantity to compute is the probability density of getting extinct, i.e. $x = 0$, after a time t and starting from an initial condition x_0, the first passage time (FPT) density [66]:

$$p_{\text{FPT}}(t|x_0) = x_0\sqrt{\frac{\theta}{2\pi t^3}}\exp\left(-\frac{\theta}{2t}\left(x_0 - t/\tau\right)^2\right).$$

Closely related is the extinction probability at time t, which consists of summing up all the events that have led to hitting the boundary from 0 to t:

$$p_{\text{extinct}}(t|x_0) = \int_0^t dt'\, p_{\text{FPT}}(t'|x_0) = 1 - p_{\text{surv}}(t|x_0),$$

where p_{surv} is the survival probability. This allows us to compute, for example, the probability for a clone having abundance in $[x_{\min}, x_{\max}]$ to go extinct:

$$f_{\text{turn}}(t|x_{\min}, x_{\max}) = \frac{\int_{x_{\min}}^{x_{\max}} dx P_{\text{stat}}(x) p_{\text{ext}}(t|x)}{\int_{x_{\min}}^{x_{\max}} dx P_{\text{stat}}(x)}. \tag{S10}$$

We can also interpret this quantity as the average fraction of trajectories that have been replaced by newly generated ones (in the stationary regime), i.e. the turnover fraction.

To move from the conditioning on x to the experimental counts n, we write down:

$$p_{\mathrm{FPT}}(t,n) = \int dx\ P_{\mathrm{stat}}(x)\ P_{\mathrm{noise}}(n|Me^x/N_{\mathrm{cell}})\ p_{FPT}(t|x),$$

and we can use it to compute the probability of a clone to get extinct at time t given the experimental observation n at the present time:

$$p_{\mathrm{surv}}(t|n) = 1 - \frac{\int_0^t\ dt' p_{\mathrm{FPT}}(t',n)}{\int\ dx\ P_{\mathrm{stat}}(x)\ P_{\mathrm{noise}}(n|Me^x/N_{\mathrm{cell}})}. \tag{S11}$$

B. Computing the likelihood with Monte Carlo sampling

To write down the likelihood, we need the probability of observing the experimental counts of the clone i in a given set of T samples according to a model parametrized with $\mathbf{p}$:

$$P(n_i^1, n_i^2, \ldots, n_i^T|\mathbf{p}) = \int \prod_{j=1}^{T} dn'^j P_{\mathrm{dyn}}(n_i'^1, n_i'^2, \ldots, n_i'^T|\mathbf{p}_{\mathrm{dyn}}) P_{\mathrm{noise}}(n_i^j|n'^j, \mathbf{p}_{\mathrm{noise}}), \tag{S12}$$

which generalizes Eq. 1 of the main text for an arbitrary number of samples. The model, in this case a GBM, sets the joint probability over the hidden variables, P_{dyn}, and the noise model, P_{noise}, defines the probability of sampling an experimental count given the true hidden count. If the analytical form of P_{dyn} is known, one can directly solve the integral with some numerical method like we did in Sec. S2. The GBM admits an analytical solution for the propagator; however we choose an alternative approach to evaluate this integral through Monte Carlo (MC) integration. A first reason for that is the fact that MC scales much better with the number of samples T with respect to a multidimensional numerical integration. Second, the next model that couples memory cells and plasmablasts is not analytically solvable but relatively easy to sample, making an MC-based approach the best choice for later.

The integral above can be evaluated with the classical expression of MC integration, where N_{MC} samples are generated from the dynamical model, and with N_i^j we indicate the MC samples of experimental count for the clone i at time j sampled from the dynamical model, $\{N_i^1, \ldots N_i^T\} \sim P_{\mathrm{dyn}}(\cdot|\mathbf{p}_{\mathrm{dyn}})$:

$$P(n_i^1, n_i^2, \ldots, n_i^T|\mathbf{p}) = \lim_{N_{\mathrm{MC}}\to\infty} \frac{1}{N_{\mathrm{MC}}} \sum_{N_i^1,\ldots,N_i^k \sim P_{\mathrm{dyn}}}^{N_{\mathrm{MC}}} \prod_{j=1}^{T} P_{\mathrm{noise}}(n_i^j|N_i^j, \mathbf{p}_{\mathrm{noise}}). \tag{S13}$$

A practical problem of using this approach is that most of the sample generated by a GBM will have small real counts (the stationary distribution is dominated by small n'), leading to many samples for small experimental counts, i.e. $n = 0, 1$, but very few samples for larger experimental counts, making the estimate of their probabilities very noisy. A typical way to overcome the problem is to use importance sampling, where we introduce an arbitrary function $g(n_i'^1, n_i'^2, \ldots, n_i'^T)$ that samples more evenly our space. The MC evaluation then becomes:

$$P(n_i^1, n_i^2, \ldots, n_i^T|\mathbf{p}) = \lim_{N_{\mathrm{MC}}\to\infty} \frac{1}{N_{\mathrm{MC}}} \sum_{N_i^1,\ldots,N_i^T \sim g}^{N_{\mathrm{MC}}} \frac{P_{\mathrm{dyn}}(N_i^1, \ldots, N_i^T)}{g(N_i^1, \ldots, N_i^T)} \prod_{j=1}^{T} P_{\mathrm{noise}}(n_i^j|N_i^j, \mathbf{p}_{\mathrm{noise}}). \tag{S14}$$

We do not know, in general, the explicit shape of the function P_{dyn}, making this expression not usable. However, in general, we know the stationary distribution P_{stat}, and we can split the joint probability as $P_{\mathrm{dyn}}(n'^1, \ldots, n'^T) = P_{\mathrm{stat}}(n'^1) P_{\mathrm{prop}}(n'^2, \ldots, n'^T|n'^1)$. The second object, P_{prop}, is the propagator of the dynamical model that can be sampled by solving numerically the stochastic differential equation. Similarly, we split the importance function into those two parts, with a first arbitrary function for sampling the initial point and the same propagator as before: $g(n'^1, \ldots, n'^T) = h(n'^1) P_{\mathrm{prop}}(n'^2, \ldots, n'^T|n'^1)$. If we rewrite the equation above by plugging in these new definitions, we have:

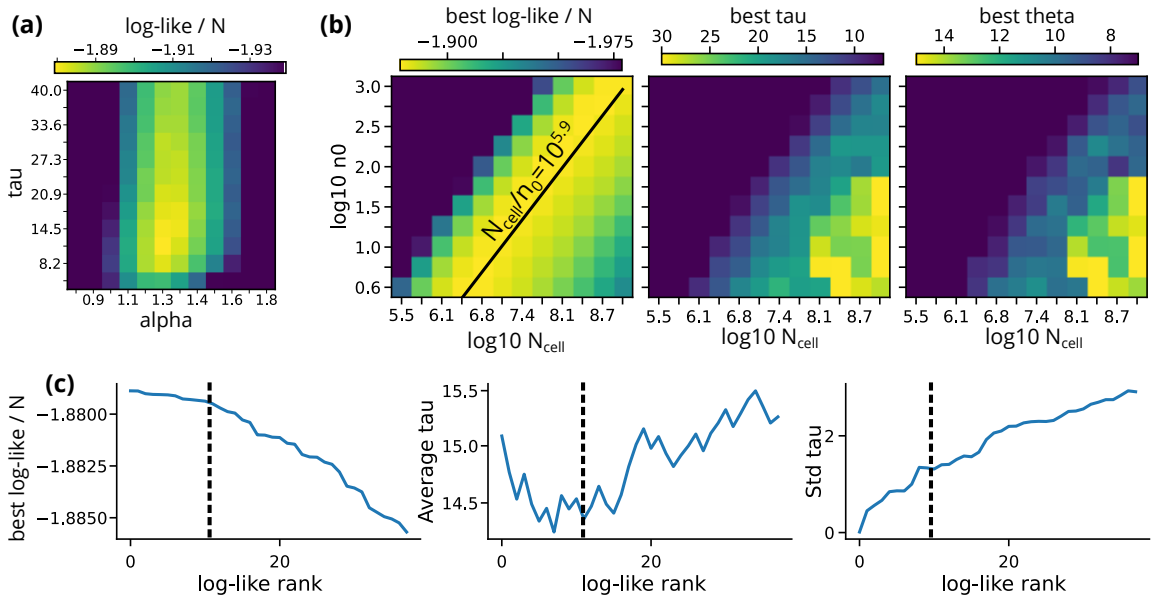
$$P(n_i^1, n_i^2, \ldots, n_i^T|\mathbf{p}) = \lim_{N_{\mathrm{MC}}\to\infty} \frac{1}{N_{\mathrm{MC}}} \sum_{N_i^1\sim h; N_i^2,\ldots,N_i^T \sim P_{\mathrm{prop}}}^{N_{\mathrm{MC}}} \frac{P_{\mathrm{stat}}(N_i^1)}{h(N_i^1)} \prod_{j=1}^{T} P_{\mathrm{noise}}(n_i^j|N_i^j, \mathbf{p}_{\mathrm{noise}}). \tag{S15}$$

In other words, the importance sampling is imposed only on the initial stationary distribution, $N_i^1 \sim h$, for which we can explicitly compute the correction factor $P_{\mathrm{stat}}(N)/h(N)$, while the rest of the temporal samples are generated using the differential equation. By choosing a proper shape of h, in our case, the variance of the likelihood can be reduced of more than one order of magnitude compared to Eq. S13.

C. Generating samples of the GBM

The GBM is defined by 4 free parameters: τ, θ, s, and n_0. We want to generate samples to compute Eq. S15 conditioned on the fact that the experimental counts at the first step are larger than zero: $P(n_i^1, n_i^2, \ldots, n_i^T | n_i^1 > 0, \mathbf{p})$. This condition allows us to exclude from the inference all the clones that are born between the first time point and the next ones, saving computational time in generating those events.

As a first step, we generate $N_{\rm MC}$ samples of log-counts at the first time point from the importance function $h(n^1)$. This function is chosen as a gamma distribution over the space of the log-counts. It depends on two parameters, the shape and the scale, which are chosen to minimize the likelihood variance. This is done by selecting a set of GBM parameters $\mathbf{p}$ reasonably close to the maximum and by computing many repetitions of the log-likelihood over which we compute the variance to minimize. Each of the $N_{\rm MC}$ samples are then used as initial conditions of a numerical integration of Eq. S6. The time step dt is chosen to be smaller than all the time scales of the equation, i.e., τ and θ, and typically of the order $0.1 - 0.01$ years. Log-counts that reach the boundary $x = 0$ are considered extinct. After the desired interval of time, the log-count values are memorized and plugged into the noise probability $P_{\rm noise}$ to compute Eq. S15. The noise probability is dataset specific and independently learned previously, as described in Sec. S2. Simulations of GBM trajectories are generated in C++ functions wrapped into Python. The computation of Eq. S2 given the generated samples, and the computation of the likelihood given those probabilities for each clone, are performed in Python.



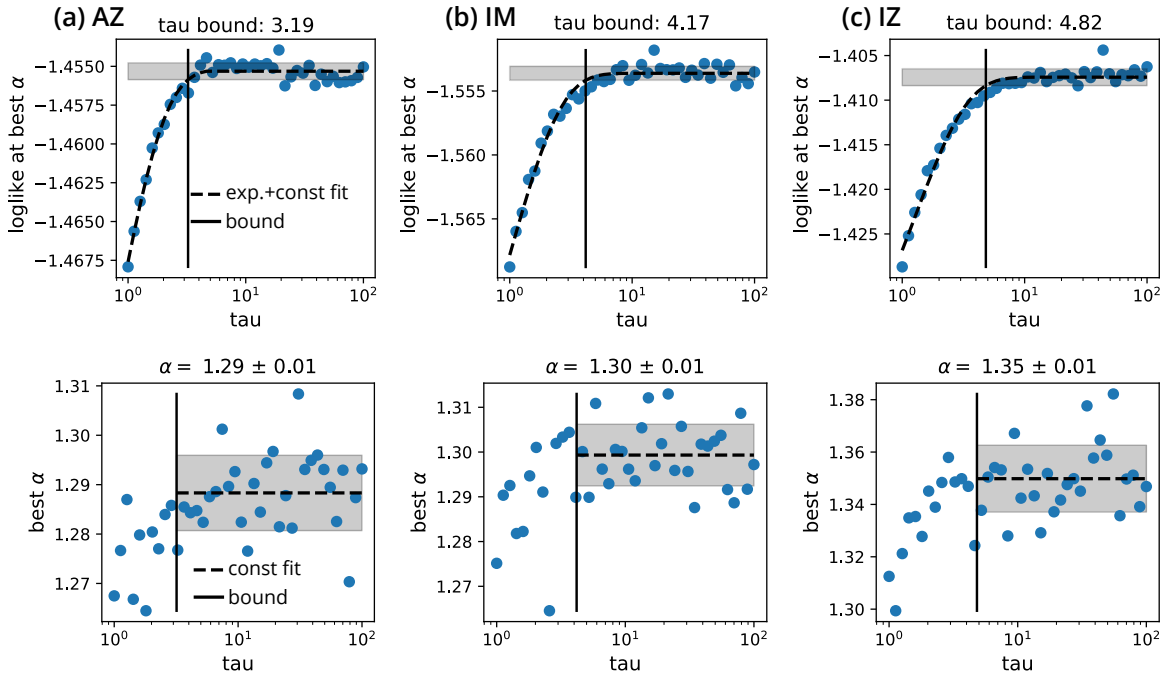
Supplementary Figure S6: **Log-likelihood shape of the GBM for Dataset A.** (a): Log-likelihood computed for the first two time points of donor 2 from *Dataset A* at fixed $n_0 = 20$, $N_{\rm cell} = 10^7$, and varying τ and $\alpha = 2\theta/\tau$. The Monte Carlo samples are $N_{\rm MC} = 2 \cdot 10^6$. (b): Best likelihood for all three time points of donor 2, maximized over τ and θ computed at different values of n_0 and M'. The first plot is the likelihood value, the second plot is the τ that maximizes it and the third one is the best value of θ. The black line fixes the ratio between $N_{\rm cell}$ and n_0 and passes at the maximum ridge. (c): Referring to the different inference procedures of panel (b) at different values of N_{cell} and n_0, we rank the trials by likelihood. We show, as a function of the best-likelihood rank, the best likelihood value, the average of best τ's across all the trials up to the rank on the x-axis, and the standard deviations of those τ's. The dashed black line is the chosen rank at which we can approximately say that the likelihood starts to drop. The parameter values and errors for this donor are the average and standard deviations of all the inference trials up to this rank.

D. Optimizing the noisy likelihood with Gaussian Processes

After the generation of MC samples, and the computation of the probability of observing a sequence of experimental counts, the likelihood can be written down under the assumption of independent clones as:

$$\mathcal{L}(\mathbf{n}^1, \ldots, \mathbf{n}^k|\mathbf{p}) = \sum_{i=1}^{N} \log P\left(n_i^1, \ldots, n_i^k \,\middle|\, n_i^1 > 0, \mathbf{p}\right) . \tag{S16}$$

In general, we fix the number of Monte Carlo samples to $N_{\mathrm{MC}} = 3 \cdot 10^5$. Since the probabilities are estimated from random samples, the likelihood itself is a noisy function that we want to maximize over the parameters $\mathbf{p}$. We usually perform the optimization in the τ-θ space or τ-α space, fixing the values of the other two free parameters of the GBM, s and n_0. An efficient and reliable way to perform this stochastic optimization in a low-dimensional space is to use Gaussian Processes to approximate the likelihood. Each evaluation of the function reduces the uncertainty of its value, and the procedure tends to focus on the areas where it is more likely that the function has a maximum. We used the function GP_OPTIMIZE of the SCIKIT-OPTIMIZE module of Python, `https://scikit-optimize.github.io/stable/index.html`. There are several hyperparameters to fix. The kernel of the Gaussian Process is chosen as a *Matérn kernel* plus a *White kernel* that encodes prior knowledge on the amplitude of the MC fluctuations. The *Matérn kernel* depends on a length scale that sets the correlation lengths, imposing a prior on the smoothness of the function along the different axes. In our case we have a very anisotropic function, Fig. S6a, which has a very long flat valley in one direction and a narrow section in the other. This motivated us to use different length scales, $50 - 100$ times larger in the τ direction than in the α direction. To balance the exploration-exploitation trade-off for sampling the function we use a *negative expected improvement* acquisition. We also set the *improvement scale*, parameter XI, to be of the order of log-likelihood differences around the maximum, i.e., $2 \cdot 10^{-4}$. A final hyper-parameter that we fix is the number of calls, N_CALLS, around 200, that is large enough to have a reliable estimate in a reasonable computational time. As a result of the optimization, we consider the average minimum and not the minimal sample, which is actually the default result that the method returns. Computationally, this can be done with the SKOPT.EXPECTED_MINIMUM function applied to the Gaussian process that has been learned at the previous step.



Supplementary Figure S7: **Log-likelihood indeterminacy for the Dataset B.** Log-likelihood shape as a function of τ_m for *Dataset B*. For the three donors with three time points, (a) AZ, (b) IM, (c) IZ, we show the log-likelihood that maximizes α at $N_{\mathrm{cell}} = 10^8$ and $n_0 = 43$, $n_0 = 58$, $n_0 = 25$ respectively (each n_0 is chosen such that the n_0/N_{cell} ratio is optimal). In the upper panels we show the minus log-likelihood that optimizes over α at fixed τ for different values of τ. The dashed line is a fit, as explained in the text. The shaded area shows the variability of points in the constant region, and the continuous line the position of the bound. The lower panels show, for the same inference procedure of the plot above, the best value of α. Here the dashed line is a fit with a constant and the shaded area the variability in that region.

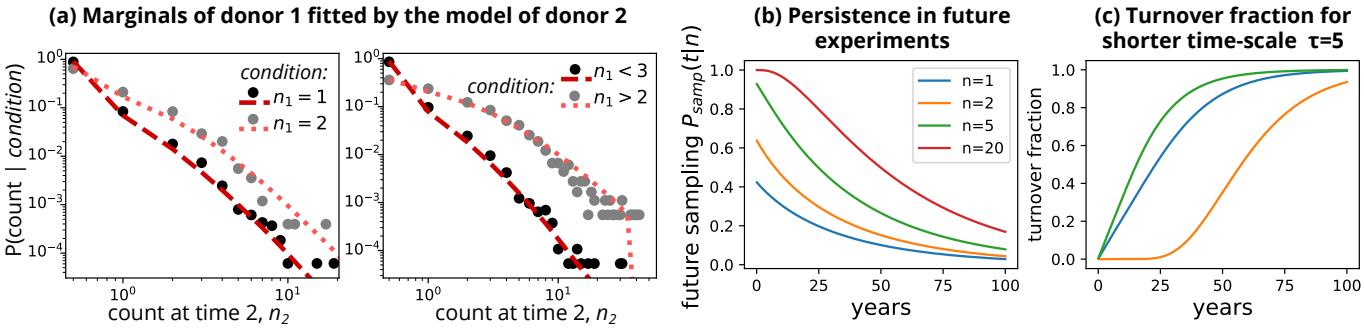
E. Inference of the GBM parameters for the single-cell dataset (*Dataset A*)

The described procedure gives us the estimate of τ_m and θ_m for the memory cell repertoire at fixed values of two other free parameters of the GBM, n_0 and s. For practical reasons it is more convenient to consider n_0 and N_{cell} instead (using Eq. S9), since s does not enter directly into the sampling procedure because of the conditioning $n^1 > 0$ that excludes the creation of new clones. N_{cell} instead is used to evaluate the noise probability, and n_0 enters the stationary distribution expression. We can run the inference for the two time points of donor 1 and for the 3 time points of donor 2 (results of Table 1). Notice that for Fig. 3f,g of the main text we use the results from the first two time points to check how well the marginal prediction generalizes on the third time point.

Since, in principle, we do not know if the other two free parameters are learnable, we repeated the inference procedure several times at different values of n_0 and N_{cell}, as shown in Fig. S6b. The likelihood shows a ridge of constant altitude that is well fitted by a fixed ratio between n_0 and N_{cell}, of order $N_{\text{cell}}/n_0 \sim 10^5$–$10^6$. More precisely, the ridge is located at $N_{\text{cell}}/n_0 = (2.38 \pm 0.01) \cdot 10^5$ for the two time points of donor 1, and at $N_{\text{cell}}/n_0 = (7.88 \pm 0.04) \cdot 10^5$ for the three time points of donor 2. The fitting procedure considers a least square method where the points are $x = \log_{10} N_{\text{cell}}$ and $y(x) = \max[\log_{10} n_0(x)]$ and the function is $y = x + c$, with $c = \log_{10}(n_0/N_{\text{cell}})$. The error derives from the covariance matrix of the least square.

The next question is how to choose the single value of τ and θ to associate to the donors among the different best parameters varying n_0 and N_{cell}. Fig. S6b shows that along the ridge those two parameters are not changing much. To better account for this variability, we employ the following procedure. We first rank the inference trials depending on the best likelihood value at different n_0 and N_{cell}. The first plot of Fig. S6c shows the best log-likelihood as a function of this rank, where we can approximately identify an initial plateau until rank ~ 10, which we can associate with the inference trials along the ridge. We can also plot the average τ within the first top r ranks as a function of r, second plot of Fig. S6c, and its standard deviation, third plot. The average and standard deviation within the $\hat{r} = 10$ top ranks, i.e., along the ridge, will be our best estimate of the parameter and its uncertainty. By varying the threshold rank $\hat{r}$, the average does not vary much, showing robustness against the arbitrary choice of $\hat{r}$. In a similar way we can compute the estimate for θ or α and its uncertainty, which are shown in Table 1.

To compute n_0 and N_{clone}, we use the estimate of the ratio N_{cell}/n_0 and Eq. S8, assuming the total number of cells to be in a range $N_{\text{cell}} = 10^{10}$–$10^{11}$ (results in Table 1). The birth rate s, Eq. S9, depends on N_{cell} only through the ratio N_{cell}/n_0. Its uncertainty is computed by propagating the error on τ_m, α and N_{cell}/n_0.



Supplementary Figure S8: **Supplementary plots to the GBM inference.** (a): Same plot of Fig 3d,e but with the data of donor 1 predicted by the model inferred from donor 2. (b): Probability of sampling a clone in a future experiment after time t having size n in the present experiment. (c): Same as Fig. 3h but for a shorter $\tau_m = 5$ and $\theta_m = 3$.

F. Inference of the GBM parameters for the cDNA amplification dataset (*Dataset B*)

For the second dataset, we considered the three donors with sufficiently deep samples at all three time points: AZ, IM, and IZ. Here we use the Negative Binomial with $b = 1$ as a noise model, which depends on the parameter a. This parameter is not learned for samples without replicates (the first time point of the three donors). Since within a given donor values of a are similar, for the sample without replicates we chose a value of a that is the average of the other two values at any given donor. This value is reported in Table 1 in the main text.

For the three donors, we found that the exponential decay time τ_m is not learnable, which can be seen in Fig. S7, upper panels. The best likelihood's dependency over α for different values of τ_m does not show a maximum but it shows a constant plateau after a steady increase. In other words, for τ_m large enough, including the case in which there is no decay at all, $\tau_m \to \infty$, we cannot see differences in the data. As discussed in the main text, this is likely due

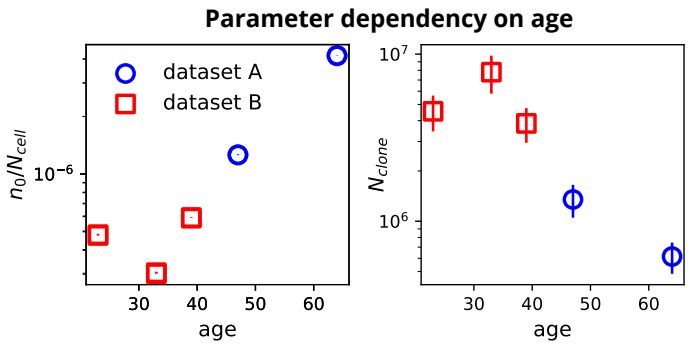
to the small temporal resolution of the dataset, where the time points are separated by 1 year, while the decay time is probably of the order of ~ 10 years, and, in that small window of time, the slow variation is not noticeable. However the value of α can be learned, because in the region of constant maximum likelihood, the best α is approximately constant (Fig. S7 lower panels).

We can find a lower bound for τ_m by searching for the point at which the likelihood starts to flatten. We fit the following piecewise function, i.e., an exponentially saturating function followed by a constant:

$$\begin{cases} A - Be^{-C\tau_m} & \text{for } \tau_m < \hat{\tau_m} \\ A - Be^{-C\hat{\tau_m}} & \text{for } \tau_m \geq \hat{\tau_m} \end{cases} \tag{S17}$$

The best fit is the dashed line in Fig. S7. Then, we compute the standard deviation σ of points in the region $\tau_m > \hat{\tau_m}$. The intersection between the constant value $A - Be^{-C\hat{\tau_m}} - \sigma$ and the exponential for $\tau_m < \hat{\tau_m}$ gives us the lower bound for τ_m, which is reported in Table 1 of the main text, and corresponds to the vertical line in the figure.

There is no minimum as a function of τ_m, but we can still study the dependency on the parameters n_0 and N_{cell} as we did in Fig. S6b. All three samples show a ridge compatible with a constant n_0/N_{cell} ratio, which we find to be $(2.11 \pm 0.03) \cdot 10^6$, $(1.70 \pm 0.01) \cdot 10^6$, and $(3.27 \pm 0.04) \cdot 10^6$ for donors AZ, IM, and IZ, respectively. From those values and assuming $N_{\text{cell}} = 10^{10}$–$10^{11}$, we can find the estimate of n_0 reported in Table 1 of the main text. However, since we do not know the precise value of τ_m, we can only find an upper bound for the introduction rate s.



Supplementary Figure S9: **Parameter dependency on age.** Normalized initial clone size, n_0/N_{cell}, and number of different clonal families N_{clone}, versus age. Errors on the left plot are typically much smaller than the point. On the right plot the errorbars are the uncertainty depending on the estimate of $N_{\text{cell}} \approx 10^{10}$–$10^{11}$.

S4. LIKELIHOOD COMPUTATION FOR THE PLASMABLAST-MEMORY DYNAMICS

A. Coupled memory-plasmablast model

The plasmablast dynamics assumes a Geometric Brownian Motion with an additional source term that models the differentiation of memory cells into plasmablasts. Given a clone i, we label its number of memory cells in the body as n'_i and its number of plasmablasts as k'_i, which follow the stochastic equation below:

$$\begin{cases} \dfrac{dn'_i}{dt} = -\dfrac{n'_i}{\tau_m} + \dfrac{n'_i}{\sqrt{\theta_m}}\eta_i(t) \\ \dfrac{dk'_i}{dt} = \rho n'_i - \dfrac{k'_i}{\tau_p} + \dfrac{k'_i}{\sqrt{\theta_p}}\xi_i(t) \end{cases}, \tag{S18}$$

where the first equation is Eq. S6 for the memory cells and $\xi_i(t)$ is a a normal Gaussian noise. The new set of parameters is ρ, the differentiation rate from memory to plasmablast, τ_p, the plasmablast average decay time, and θ_p, the time scale of the random fluctuations. Unlike the simple GBM, the system is not analytically solvable anymore, but we can derive a useful relation between the average values of n' and k' at the stationary state:

$$\langle k'_i \rangle = \rho\tau_p \langle n'_i \rangle, \qquad K'_m = \rho\tau_p N_{\text{cell}}, \tag{S19}$$

where the second equation is a consequence of the first one and connects the total number of plasmablasts generated by memory-cell differentiation, K'_m, to the number of memory cells, N_{cell}.

Inferring all three plasmablast parameters from sampled repertoires is mathematically impossible. This can be seen by writing the second equation for the frequencies: $f' = n'/N_{\text{cell}}$ and $g' = m'/K'_m$, and using the relation between numbers of cells (Eq. S19):

$$\frac{dg'_i(t)}{dt} = \rho f'_i \frac{N_{\text{cell}}}{K'_m} - \frac{g'_i}{\tau_p} + \frac{g'_i}{\sqrt{\theta_p}}\xi_i(t) = \frac{f'_i - g'_i}{\tau_p} + \frac{g'_i}{\sqrt{\theta_p}}\xi_i(t). \tag{S20}$$

The dependency on ρ disappears, and since experimental observations from samples depend only on frequencies and not on absolute counts, this parameter is not learnable. To overcome this problem, we can use empirical estimates that connect the total number of the two groups of cells. Define K' as the total number of plasmablasts. $K'_m = \phi K'$, the number of memory-derived plasmablasts, is only a fraction ϕ of the total, since plasmablasts can also be generated by naive B cells. We also consider the ratio between the total number of memory cells and the total number of plasmablasts $r = K'/N_{\text{cell}}$. The order of magnitude of these two numbers is known, and, in combination with Eq. S19, allows us to close the system of equations:

$$K'_m = \phi r N_{\text{cell}}, \quad K'_m = \rho \tau_p N_{\text{cell}}, \quad \rightarrow \quad \rho = \frac{\phi r}{\tau_p}. \tag{S21}$$

Therefore, the strategy for the inference is to assume an empirically known value for ϕ and r and to optimize the likelihood over τ_p and θ_p, which set a value for ρ through the equation above.

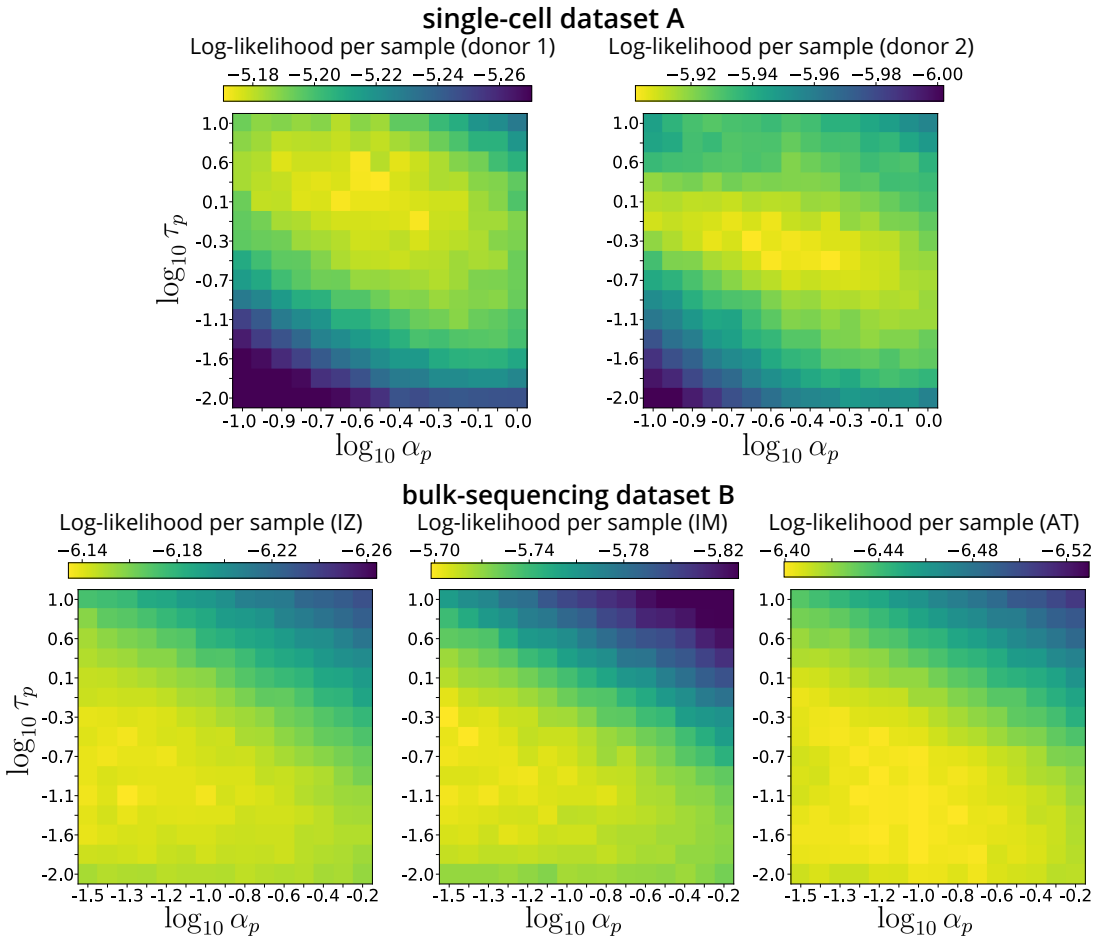
B. Sampling the stochastic equation

To compute the likelihood, we need the probability of observing experimental counts given the parameters $\mathbf{p}_{\text{dyn}} = (\tau_p, \theta_p)$, which has the shape of Eq. S12. Now we have some experimental samples of memory cells at different times t, n_i^t, and samples of plasmablasts k_i^t, and for each clone i, we want to compute $P(n_i^1, n_i^2, \ldots, k_i^1, k_i^2, \ldots | \mathbf{p})$. Each sample has a noise model learned in Sec. 4A of the main text, P_{noise}, and the dynamical model P_{dyn}, Eq. S18, governs the dynamics of the real counts n' and k'. Notice that P_{noise} is defined by the average μ, which is the frequency of the clone in the organism multiplied by total number of counts, which we call M_p for plasmablasts. Sequencing experiments contain all types of plasmablasts, not only the ones derived from memory cells, and therefore the frequency has to be computed by using K': $\mu_i = M_p k'_i / K'$ for clone i.

An analytical solution for the propagator of Eq. S18 is not possible, and therefore we used the Monte Carlo integration to compute Eq. S12 as we did for memory samples. We can also use the trick of importance sampling by choosing a proper distribution over the memory samples at the first time, g, and modify the summation of Monte Carlo samples as in Eq. S15. In the previous case we were conditioning on having the first experimental count $n_i^1 > 0$, which was allowing us to neglect the dynamics of newly generated clones. This condition, however, reduces the number of experimental samples, which is already not large for plasmablasts. Therefore, we relax this constraint but we still impose the clone to be present in the memory samples, $\sum_j n_i^j > 0$, in such a way that we consider only plasmablasts that have been generated by memory cells. As a consequence, in the MC simulation we need to generate new clones, which do not benefit from the importance sampling trick as before, since they are not sampled from the initial stationary distribution. Putting these considerations together, we perform the Monte Carlo integration as a sum of two parts: one that estimates the probability conditioned on $n^1 > 0$ with importance sampling, and the second conditioned to $n^1 = 0$ without importance sampling. The counterpart of Eq. S15 considers samples of memory-cell counts N_i^t and of plasmablasts K_i^t, and it reads:

$$\begin{aligned} P(n_i^1, \ldots, k_i^1, \ldots | \mathbf{p}) = \lim_{N_{\text{MC}} \to \infty} \frac{1}{N_{\text{MC}}} \Bigg[& \sum_{N_i^1 \sim h; N_i^2, K_i^2 \ldots \sim P_{\text{prop}}}^{cN_{\text{MC}}} \frac{P_{\text{stat}}(N_i^1)}{h(N_i^1)} \prod_j P_{\text{noise}}(n_i^j | N_i^j, \mathbf{p}) \prod_l P_{\text{noise}}(k_i^l | K_i^j, \mathbf{p}) + \\ & + \sum_{N_i^1 = 0; N_i^2, , K_i^2 \ldots \sim P_{new}}^{(1-c)N_{\text{MC}}} \prod_{j=1}^{k} P_{\text{noise}}(n_i^j | N_i^j, \mathbf{p}) \prod_l P_{\text{noise}}(k_i^l | K_i^j, \mathbf{p}) \Bigg] . \end{aligned} \tag{S22}$$

The first summation samples $c \times N_{\text{MC}}$ Monte Carlo clones (the fraction $c \leq 1$ will be discussed later) that are present at the first step, using the importance trick, and we propagated them by simulating Eq. S18 with the initial condition for the plasmablasts given by Eq. S19. The second summation considers non-existing clones at time 1, $N_i^1 = 0$, and it runs over the $(1-c)N_{\text{MC}}$ generated new clones with a process labeled P_{new}. Each clone appears at a random time uniformly sampled in the interval between the first and last experiment (separated Δt). Once created, the new



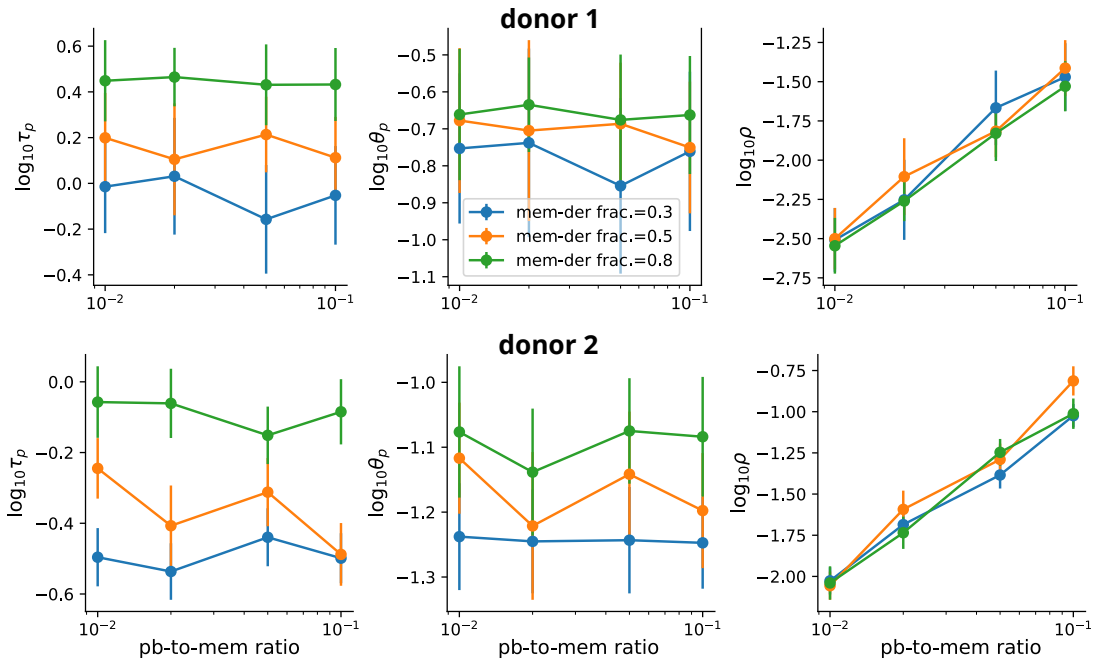
Supplementary Figure S10: **Log-likelihood shape for the memory-plasmablast dynamics.** Log-likelihood computed by varying the two optimization parameters of the memory-plasmablast dynamics τ_p and $\alpha_p = 2\theta_p/\tau_p$. The number of Monte Carlo samples is $c \times N_{\rm MC} = 5 \cdot 10^5$, where c is the fraction of samples that exists at time 1 (see SI text). The parameters of the memory cell dynamics are fixed by the previous inference on memory cells. In the case of the second dataset for which we have only a lower bound on τ_m, we chose $\tau_m = 5$ years, which is larger but close to the bound. The value of n_0 used in the simulations corresponds to the choice on $N_{\rm cell} = 10^{10}$. The plasmablast-memory ratio is chosen as $r = 0.1$, and the fraction of plasmablasts having a memory origin is $\phi = 0.5$. We show the log-likelihood for five experimental samples: donor 1 and 2 of *Dataset A* and three donors (IZ, IM, AT) of *Dataset B*.

clone has n_0 counts of memory cells and 1 plasmablast, and from those initial conditions they follow Eq. S18. We checked that for different initial conditions on the number of plasmablasts the inference result does not change. The fraction c can be found by imposing that the number of different clones is constant in time and equal to the initial count $cN_{\rm MC}$. By using Eq. S8, we obtain the creation rate $s = cN_{\rm MC}/(\tau_m \log n_0)$, and the number of new clones is $s\Delta t = (1-c)N_{\rm MC}$, from which we get $c = 1/[1 + \Delta t/(\tau_m \log n_0)]$.

The numerical simulation of Eq. S18 is performed as for the GBM case of memory cells. One additional choice to make is when a plasmablast clone goes extinct, i.e., $k_i < 1$, but the memory cells of the same clone are not yet extinct. We choose to reintroduce the plasmablast at count 1 after a time $t_{\rm new}$ given by the condition $\int_0^{t_{\rm new}} dt' n_i(t')\rho = 1$, that is, the time it takes on average for the memory cells to create 1 plasmablast. One important technical point regards the choice of the discretization temporal step dt in solving numerically the differential equation. Unlike the previous case, here the time scales for the plasmablasts are smaller and we have to be careful of always choosing dt smaller than the smallest time scale of the process. This, in addition with a slightly more complex dynamics, makes the computational time of this second case slower than the previous one.

The likelihood maximization has been performed in the space $\log_{10}\tau_p$ and $\log_{10}\alpha_p = \log_{10}(2\theta_p/\tau_p)$. We transform the variables with a logarithm because we do not know *a priori* the orders of magnitude of these quantities. The maximization was performed by using a Gaussian Process to smoothen the noisy landscape of the Monte Carlo estimated likelihood. Unlike for memory cells, we do not impose different length-scale hyper-parameters in the kernel

of the Gaussian Process. We found that, when learnable, the log-likelihood landscape is not as anisotropic as in the memory cell case (see next Fig. S10a). We compute errors on the obtained optimal values by computing the Hessian of the average Gaussian Process function, which approximates the log-likelihood landscape in the absence of Monte Carlo noise. From the Hessian H, the error of parameter p_i is $\sigma_i = \sqrt{H^{-1}_{i,i}}$.



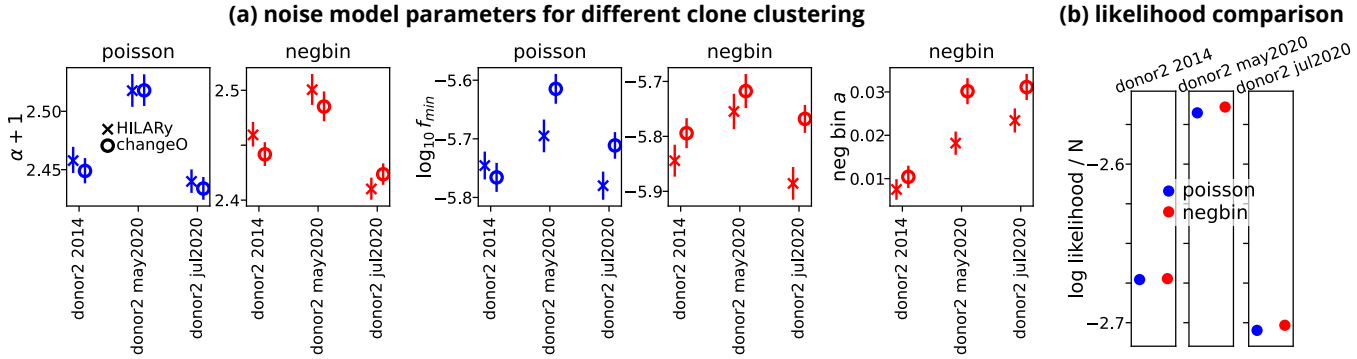
Supplementary Figure S11: **Inferred parameters of the memory-plasmablast dynamics.** The parameters are computed across a range of different plasmablast-to-memory cell ratio, r, on the x-axis, for different values of the fraction of plasmablasts that derive from memory cells, ϕ (colors). The error bars are computed using the inverse of the likelihood Hessian, where the Hessian is obtained from the smoothed approximation of the likelihood by the Gaussian Process that is used to maximize the noisy function.

C. Likelihood optimization for *Dataset A*

We perform the inference procedure for both donors. For the first donor we exclude the first plasmablast sample given its small depth and we use the remaining two memory and two plasmablast samples. For the second donor we run the inference by using all three samples of memory cells and the second and third ones of plasmablasts, excluding the first sample for reasons of sequencing depth, and the last sample for reasons of performance. However, we checked that including it does not change the results. The inference uses the memory cell parameters learned at the previous step and n_0 fixed by the choice $N_{\text{cell}} = 10^{10}$. This leads to the likelihood landscape shown in Fig. S10. It shows a clear maximum, which still remains when varying the empirical parameters r and ϕ. The results for different learning trials at different $r = 0.01$–0.1 and $\phi = 0.3$–0.8 are shown in Fig. S11, where τ_p and θ_p are learned and ρ is obtained from the constraint S21.

D. Likelihood optimization for *Dataset B*

The longitudinal structure of this dataset, with relatively short intervals between timepoints, makes it suitable to learn the fast plasmablast dynamics. However, we obtain a flat likelihood surface that keeps going for increasingly smaller values of α_p. This flatness does not allow us to learn the parameters. We suspect that the reason is the noise in the experimental plasmablast samples, for which we have learned a Negative Binomial model with a much larger variance than a Poisson model (see Sec. S2). The average decay of plasmablasts and the relation with memory cells is hidden in the noise. Consistently, the best likelihood area includes region of maximal noise $\alpha_p \to 0$, i.e., $\theta_p \to 0$.



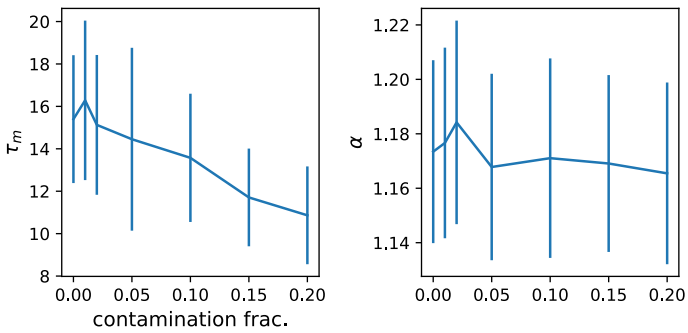
Supplementary Figure S12: **Comparison of the noise parameters inferred using HILARy and ChangeO inferred families.** Poisson (blue) and Negative binomial (red) model parameters inferred from the memory B-cell samples grouped into clones using two different clustering algorithms, HILARy and ChangeO (shown with different point markers), for the second donor of *Dataset A*. (a) model parameters computed as described in Sec. S2. (b) log-likelihood per sample for the ChangeO families.

	decay time, τ_m (year)	fluct. time θ_m (year)	power law exponent, α	initial clone size, n_0	rate of new clones, s (1/year)	number of clones, N_{clone}	diff. rate ρ (1/year)	decay time τ_p (year)	fluctuation time θ_p (year)
Hilary	14.5 ± 1.3	8.7 ± 0.7	1.20 ± 0.04	10^4–10^5	$(9.0 \pm 1.6) \cdot 10^3$	$(1.2$–$1.5) \cdot 10^6$	0.01–0.1	0.25–1	0.05–0.1
ChangeO	13.0 ± 1.0	7.6 ± 0.4	1.17 ± 0.04	10^4–10^5	$(8.3 \pm 1.7) \cdot 10^3$	$(1.0$–$1.3) \cdot 10^6$	0.01–0.1	0.3–1	0.05–0.09

Supplementary Table S1: **Comparison of the parameters of the dynamical models inferred using HILARy and ChangeO inferred families.** Results are presented for donor 2 of *Dataset A* for both the memory B cell and plasmablasts models. Note that the first line (HILARy) corresponds to the numbers reported in Tables 1 and 2.

	decay time, τ_m, years	fluctuation time, θ_m
donor 1 (A)	8.8 ± 1.8	6.5 ± 1.6
donor 2 (A)	9.6 ± 2.4	6.5 ± 1.5

Supplementary Table S2: Inferred parameters of the Geometric Brownian Motion for the memory B-cell dynamics obtained by defining a clone as a set of identical sequences, for donor 2 of *Dataset A*.



Supplementary Figure S13: **Model inference with contamination of plasmablasts in the memory samples.** We repeated the memory-cell GMB inference by substituting a fraction (x axis) of memory cells with plasmablasts taken at the same time, for donor 2 of *Dataset A*. The inference fixes $N_{\text{cell}} = 10^{10}$ and $n_0 = 12700$ (a point along the ridge of maximum likelihood of Fig. S6) and finds the best τ_m and α by using the Gaussian Process method described in Sec. S3. The error is computed with the inverse Hessian of the log-likelihood approximated with the Gaussian process. Notice that, for fraction 0, the precise values and errors are slightly different than in Tab. 1 because the reported parameters are averaged along the $N_{\text{cell}} - n_0$ ridge as explained in Sec S3.