Prepared for review, July 2023

SURROGATE SELECTION OVERSAMPLES EXPANDED T CELL CLONOTYPES

BY PENG YU[1,*], YUMIN LIAN[2], CINDY L. ZULEGER[3,4], RICHARD J. ALBERTINI[5], MARK R. ALBERTINI[3,4,6,†] AND MICHAEL A. NEWTON[1,4,7,‡]

[1]*Department of Statistics, University of Wisconsin, Madison, *peng.yu@wisc.edu*

[2]*Department of Chemistry, Laboratory of Genetics, University of Wisconsin, Madison*

[3]*Department of Medicine, School of Medicine and Public Health, University of Wisconsin, Madison*

[4]*Carbone Cancer Center, University of Wisconsin, Madison*

[5]*University of Vermont, Burlington, VT, USA*

[6]*Medical Service, William S. Middleton Memorial Veterans Hospital, Madison, †mralbert@wisc.edu*

[7]*Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, ‡manewton@wisc.edu*

Inference from immunological data on cells in the adaptive immune system may benefit from modeling specifications that describe variation in the sizes of various clonal sub-populations. We develop one such specification in order to quantify the effects of surrogate selection assays, which we confirm may lead to an enrichment for amplified, potentially disease-relevant T cell clones. Our specification couples within-clonotype birth-death processes with an exchangeable model across clonotypes. Beyond enrichment questions about the surrogate selection design, our framework enables a study of sampling properties of elementary sample diversity statistics; it also points to new statistics that may usefully measure the burden of somatic genomic alterations associated with clonal expansion. We examine statistical properties of immunological samples governed by the coupled model specification, and we illustrate calculations in surrogate selection studies of melanoma and in single-cell genomic studies of T cell repertoires.

Funding. This research was supported in part by the National Science Foundation (grant 2023239-DMS), and by grants from the National Institutes of Health: R01 GM102756, P01 CA022443, P01 CA250972, P50 CA278595, UL1 TR002373, P50 CA269011, and P30 CA014520. This work was also supported by resources at the William S. Middleton Memorial Veterans Hospital, Madison, WI, USA, and the UW Carbone Comprehensive Cancer Center. Additional support was provided by Ann's Hope Foundation, Taking on Melanoma, the Tim Eagle Memorial, and the Jay Van Sloan Memorial from the Steve Leuthold Family Foundation, philanthropic support in the USA. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or the views of the Dept. of Veterans Affairs or the United States Government.

1. Introduction.

1.1. *Overview.* With thymic-derived lymphocytes (i.e., T cells) sampled from peripheral blood or some other tissue compartment (e.g., tumor-infiltrating lymphocytes), any techniques that would enrich the sample for disease-relevant cells could be useful, considering the complexity of a typical T cell population and the potential for an improved understanding

Keywords and phrases: Bayes's rule, clonal expansion, diversity statistic, enrichment, exchangeable birth-death processes, experimental design, single cell sequencing, size bias, somatic mutation, Yule-Simon law.

of the immune response to disease. For example, at writing we have no effective biomarkers to predict how a melanoma patient will respond to immune checkpoint inhibition therapy, though responses among similar patients may vary from morbid toxicity to full recovery (e.g., Ganesan and Mehnert, 2020; Shum, Larkin and Turajlic, 2022).

Surrogate selection restricts a lymphocyte sample *in vitro* to cells whose somatic ancestors had acquired and thus transmitted to them specific, selectable mutations. Selection assays based on mutations of the hypoxanthine-guanine phosphoribosyltransferase (HPRT) gene are most well studied, though the approach applies to any mutations that are neutral with respect to the immune response (Kaitz et al., 2022). As an immune-system probe, HPRT surrogate selection has been used to study a variety of environmental effects and disease processes (Albertini, Castle and Borcherding, 1982; Albertini, 2001; Kaitz et al., 2022). With continued focus on disease studies, we examine the sampling effects of surrogate selection; selected cells may represent *in vivo* amplified clones that are more likely to be disease relevant than clones of randomly sampled cells, and we seek a more thorough understanding of this enrichment phenomenon for the sake of improved experimental design and data analysis.

The idea that surrogate selection can enrich for clonally amplified T cells has provided a rationale in many studies, though quantitative treatments of this experimental-design strategy remain very limited. Statistical procedures have been deployed to test from sequence data the null hypothesis that enrichment is absent, and the mounting evidence supports the alternative (e.g., Pei et al., 2014; Zuleger et al., 2020). Considering cell growth dynamics, one would predict an increased prevalence of various somatic mutations in cells within an actively proliferating clone compared to a relatively quiescent one. Then conditioning on the presence of some such mutation in a sampled cell, Bayes's rule would imply that the cell is more likely to be from the proliferating than the quiescent clone. Surrogate selection thus relies on the biological consequences of *in vivo* clonal proliferation to enrich for activated T cells in individuals with ongoing immunological response to disease. Understanding this enrichment effect is complicated by the enormous complexity of T cell population and properties of the distribution of clone sizes, but resolving these complications will inform investigations of surrogate selection as a mechanistic probe for fundamental biological/immunological processes. The main contribution of the present work is to quantify the enrichment effect of surrogate selection in an idealized but structurally relevant setting, and to leverage basic stochastic-process theory to confirm and characterize the enrichment phenomenon in this model. Our formulation also enables a study of distributional properties of elementary diversity statistics, of the type often used in experimental studies. We show that samples identified using surrogate selection have lower expected sample diversity, in agreement with empirical studies.

Our theoretical analysis exposes an interesting statistical prediction concerning somatic mutations that are unrelated to any selection assay. From contemporary single-cell genomic studies, we associate T cell clone sizes with estimates of somatic mutation burden, and thereby provide a new measure of somatic burden of a T cell receptor.

1.2. *Immunological setting.* Consider a person's T cell repertoire, comprised of perhaps 10^{11} or more CD4+ and CD8+ naive, effector, and memory T cells, and partitioned into clonotypes within each of which the T cell receptor (TCR) sequence of the cells is constant (e.g., Nikolich-Žugich, Slifka and Messaoudi, 2004; Pennock et al., 2013; van den Broek, Borghans and van Wijk, 2018). The number of T cells in each clonotype fluctuates over time and usefully may be viewed as a stochastic process (Currie et al., 2012; Hodgkin, Dowling and Duffy, 2014; Desponds, Mora and Walczak, 2016; Gaimann et al., 2020; Smith et al., 2020; Molina-París and Lythe, 2021). Notably, a T cell receptor's cognate antigen may induce cell division and expansion of the associated clonotype when appropriate costimulatory molecules are present. Complexity of the adaptive immune response warrants highly

detailed stochastic-model dynamics, perhaps accounting for clonal competition or adaptation (e.g., Stirk, Molina-París and van den Berg, 2008; Lythe and Molina-París, 2018; Rane et al., 2018; Duque et al., 2020). However, even structurally simple models can support certain lines of investigation and can guide statistical analysis in the growing number of empirical studies. T cell receptor repertoire analysis has been critical in studies investigating antitumor responses as well as immune-related toxicity following treatment with immune-checkpoint blockade (e.g., Fairfax et al., 2020; Valpione et al., 2020; Lozano et al., 2022; Valpione et al., 2021).

1.3. *Surrogate selection.* In the absence of an assay to measure the proliferation history of a sampled T cell, surrogate selection provides an indirect measurement through the lens of neutral somatic mutation. The most well-studied case leverages an assay to score somatic mutations of hypoxanthine-guanine phosphoribosyltransferase (HPRT) (Albertini et al., 1990; Albertini, 2001). Other assays rely on an efficient approach to screen mutations in phosphoinositolglycan class A (PIG-A) genes (Peruzzi et al., 2010; Dobrovolsky et al., 2017). Coding an enzyme within the purine salvage pathway, HPRT normally helps to recycle nucleotide bases from degraded DNA. Its post-translational modifications also confer cytotoxicity to purine analogs, including 6-thioguanine (6TG). Cultured lymphocytes are thus unable to grow in the presence of 6TG unless they have incurred an inactivating HPRT mutation. Each surviving T cell in an HPRT assay reports that an HPRT mutation occurred in that T cell or in one of its somatic ancestors. The assay has been used to monitor somatic mutations in many settings, including, for example, in Chernobyl liquidators (Jones et al., 2002), in Iraq war veterans (Nicklas et al., 2015), and in studies of environmental exposures. Kaitz et al. (2022) reviews the implicit model for surrogate selection and the literature using HPRT surrogate selection in autoimmune diseases, cardiac transplantation, infectious diseases, a hematological disease, and cancer.

1.4. *Summary of findings.* The rationale for surrogate selection in disease studies is that it provides an enrichment for relevant T cell clonotypes. Some care is required in this argument, since while a large, expanded clonotype has higher sampling probability than any smaller clonotype, the vast diversity within a typical T-cell repertoire means that even large clonotypes remain a small fraction of the total population; indeed, most sampled cells come from small clonotypes. Basic stochastic process theory guides our effort to balance these factors. We find that if at any time point the vector of clonotype sizes in a repertoire is exchangeable, and if the temporal development of any one clonotype follows a sufficiently regular birth-death process, then surrogate selection via neutral somatic mutation enriches the sampled cells for those of larger clonotypes. We examine the impact of surrogate-selection on the expected value of sample diversity statistics. In empirical validations, we re-examine single-cell data from publicly available T cell repertoire samples that were obtained via 10x Genomics sequencing; in doing so we compute cell-level somatic burden statistics and associate this burden with clonotype size. We also review sample diversity statistics from available surrogate-selection studies.

2. One developing clonotype.

2.1. *Model set up.* Our calculations begin by considering one clonotype of the many within an individual subject's T cell repertoire. For definiteness, we label this clonotype σ, recognizing that σ resides in a large finite label set $\mathcal{S}$, which we associate with the set of possible T cell receptor sequences. At time $t \geq 0$ relative to some reference time point $t = 0$ (e.g., birth), clonotype σ consists of $N_\sigma(t)$ cells. If clonotype σ is ever non-empty, then

there is some origin time, say τ_σ, such that $N_\sigma(t) = 0$ for $t < \tau_\sigma$ and $N_\sigma(t) > 0$ only at times $t \geq \tau_\sigma$. We suppose that $N_\sigma(\tau_\sigma) = 1$; that is, the clonotype originates upon successful completion of receptor-forming recombination events (Elhanati et al., 2018). After positive and negative selection induce thymocyte maturation, clonotype cells egress from the thymus and distribute themselves throughout the body; we expect this all occurs on a short time scale compared to the timing of typical observations, which might be from a mature subject's peripheral blood or tumor-infiltrating lymphocytes, for example.

The stochastic process $\{N_\sigma(t) : t \geq 0\}$ fluctuates in response to all sorts of cell-biological factors affecting cells in the clonotype, and must reflect a complex birth-death process (e.g., den Braber et al., 2012; Desponds, Mora and Walczak, 2016; Zhan et al., 2017). For example, in the presence of appropriate cytokines, T cell receptor interaction with cognate antigen triggers cell proliferation, while apoptotic signals can induce cell death. Our understanding of repertoire maintenance further supports the notion that if $N_\sigma(s) = 0$ at time $s > \tau_\sigma$, then $N_\sigma(t) = 0$ for all $t \geq s$. This is analogous to the infinite-alleles assumption in population genetics; here it means that a clonotype can only emerge once.

2.2. *The branching tree.* Following clonotype σ over time from τ_σ, there is a series of event times at which cells in the clonotype either divide or die. Were we able to trace σ's complete history, we would record a binary tree, such as in Figure 1. At some observation

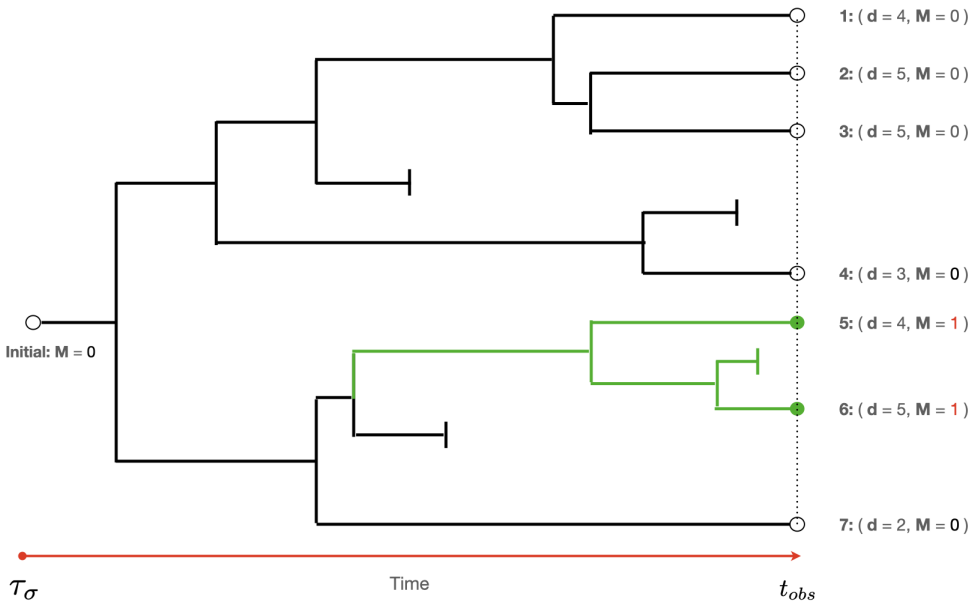

FIG 1. *Binary tree formed by a developing clonotype, showing examples of cell division, cell death and mutation, and noting the number d of cell divisions experienced by each extant cell at time* t_{obs}. *Green circles (extant cells 5 and 6) denote mutant T cells. Empty circles (1, 2, 3, 4 and 7) denote wild type T cells. Green lines denote evolution of mutant cells. Short vertical lines denote cell death.*

time t_{obs}, each leaf of the tree is an extant cell that has experienced a number of cell divisions since τ_σ. This division number is also called the depth of the leaf node. For a cell randomly sampled from the clonotype, let D_σ denote this division number; it has a probability distribution induced both by the stochastic development of σ and by the random selection of the extant cell. Fortunately, this distribution has been the subject of extensive study in the

context of random binary trees (e.g., Lynch, 1965; Mahmoud, 1992; Aldous, 1996; Steel and McKenzie, 2001; Mahmoud and Neininger, 2003).

In the Yule model for trees, each cell division acts on a random cell, as if by a pure-birth process without cell death. This symmetry over cell identity allows various explicit computations. In fact, the probability generating function (p.g.f.) of D_σ is

$$G_n(z) = E\left\{ z^{D_\sigma} \,\middle|\, N_\sigma(t_{\text{obs}}) = n \right\} = \frac{\langle 2z \rangle_{n-1}}{n!}, \tag{1}$$

which is the formulation presented in (Mahmoud, 1992, Page 71-74), Eq. (2.4).[1] Here $\langle x \rangle_n = x(x+1)(x+2)\cdots(x+n-1)$ is the rising factorial, which is conveniently expressed in terms of Gamma and Beta functions Γ and B as:

$$\frac{\langle x \rangle_{n-1}}{n!} = \frac{\Gamma(x+n-1)}{\Gamma(x)\Gamma(n+1)} = \frac{1}{(x+n)(x+n-1)} \cdot \frac{1}{B(x, n+1)}.$$

The p.g.f. G_n helps us connect the T cell repertoire with surrogate-selection dynamics. Before pursuing that calculation, we note that the expectation and variance of D_σ are also available, with both well approximated by twice the natural logarithm of n, and that as n increases, $\{D_\sigma - 2\log(n)\} / \sqrt{2\log(n)}$ converges in distribution to a standard normal variate (Brown and Shubert, 1984; Mahmoud and Neininger, 2003). Roughly, a randomly sampled cell from a randomly proliferating clonotype of current size n (and ignoring cell death) has experienced about $2\log(n)$ cell divisions since receptor formation in the thymus. Sampling from the conditional distribution of $D_\sigma | N_\sigma(t_{\text{obs}}) = n$ is reported in Figure 2, revealing this proliferation effect for a handful of clonotype sizes. For completeness, we note the p.m.f. of D_σ is, as derived in Lynch (1965),

$$P\{D_\sigma = d \mid N_\sigma(t_{\text{obs}}) = n\} = \frac{2^d}{n!} S(n-1, d), \qquad d = 0, 1, \cdots, n-1, \tag{2}$$

where $S(n-1, d)$ is the unsigned Stirling number of first kind.

2.3. *Neutral mutations.* Surrogate selection aims to use neutral genomic mutations – mutations that do not affect clonotype growth dynamics – as probes to report on these very same dynamics. Uncorrected mitotic errors or other mutagenic effects are expected to occur at some rate throughout the developing repertoire. We focus on mitotic mutations that affect a single daughter cell, that are irreversible, and that occur independently across cell divisions. Less prevalent mechanisms may induce mutations in both daughter cells (e.g., double-stranded breaks) or separately from mitosis (e.g., ionizing radiation), and statistical formulations may be adapted to these cases (e.g., Kendall, 1960; Roshan, Jones and Greenman, 2014). We use $\theta \in (0, 1/2)$ to denote the relative frequency of mutations at a given locus (e.g., HPRT) per daughter cell; i.e., 2θ is the mutation frequency per cell division.

Consider the thought experiment to sample a single cell uniformly at random from the extant clonotype σ at time t_{obs}, and let M_σ be the binary $(0/1)$ indicator that the sampled cell harbors a mutation at the locus in question. We recognize that M_σ really indicates that a mutation event occurred somewhere in the ancestral lineage of the cell, and thus

$$P\{M_\sigma = 1 | D_\sigma = d, N_\sigma(t_{\text{obs}}) = n\} = 1 - (1-\theta)^d \tag{3}$$

[1] In Mahmoud (1992), a binary tree is assumed to contain n internal nodes, and Eq. (2.4) cares about the $n+1$ external nodes (leaves) of the corresponding extended binary tree. In Steel and McKenzie (2001), following Mahmoud (1992), the Yule tree is said to contain $n+1$ leaves. Our notation is slightly different as we use n to denote leaf numbers. In our setting $n \geq 1$ and $D_\sigma \geq 0$.

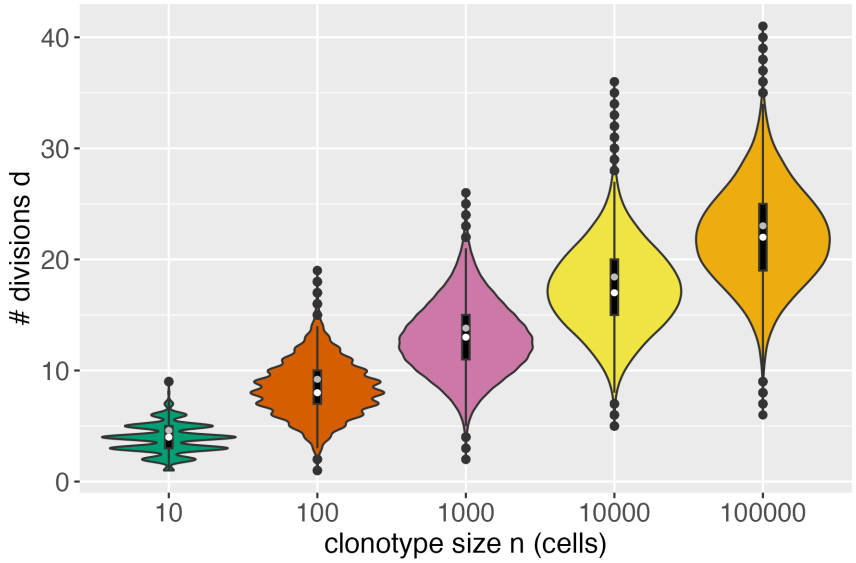

FIG 2. *Proliferation effect: Shown are violin plots of the division number D_σ for cells in randomly developed binary trees, having various sizes, n, at observation time. We used* **R** *packages* ***ape****, to simulate Yule trees, and* ***adephylo****, to count divisions (Paradis and Schliep, 2019; Jombart, Balloux and Dray, 2010). Each plot summarizes 100,000 simulated D_σ values. Empirical medians (white) and asymptotic means $2\log(n)$ (grey) are shown.*

where D_σ is the division number for this random cell. (The cell is not mutant if none of the d opportunities for mutation yield such.) Incidentally, (3) implies that M_σ and $N_\sigma(t_{\text{obs}})$ are conditionally independent given D_σ. Our first finding concerns the rate of mutant genotype in clonotypes of a given size, and is obtained by marginalizing the distribution of D_σ. With neutral mutations in a Yule tree model, define $\psi_n := P\{M_\sigma = 1 | N_\sigma(t_{\text{obs}}) = n\}$, and note,

$$\begin{aligned}
\psi_n &= \sum_{d=0}^{\infty} P\left(M_\sigma = 1 | D_\sigma = d\right) P\left\{D_\sigma = d | N_\sigma(t_{\text{obs}}) = n\right\} \\
&= \sum_{d=0}^{\infty} \left\{1 - (1-\theta)^d\right\} P\left\{D_\sigma = d | N_\sigma(t_{\text{obs}}) = n\right\} \\
&= 1 - G_n(1-\theta) \\
&= 1 - \frac{\Gamma(n+1-2\theta)}{\Gamma(n+1)\Gamma(2-2\theta)} \approx 1 - \frac{1}{n^{2\theta}\,\Gamma(2-2\theta)},
\end{aligned} \tag{4}$$

with the approximation on the last line improving for increasing n. Result (4) quantifies the intuition that proliferating clonotypes provide a greater number of chances for mutation. With $\theta > 0$, $\lim_{n\to\infty} \psi_n = 1$, and so an ever-proliferating clonotype is eventually dominated by mutant cells. This matches limit theory for birth-death processes in which the growth rate of mutant cells is no less than that of wild-type cells (e.g., Cheek and Antal, 2018).

We are not too concerned with the total number of mutant cells in the clonotype, whose expected value is n time the per cell rate in (4), though our diversity calculations in Section 3.5 rely on this distribution. That total mutant count is interesting in other settings, and is governed by the Luria-Delbrück distribution; see Angerer (2001) or Roshan, Jones and Greenman (2014) for the exact, non-asymptotic formulation. The reader may check that our formula (4) matches the first-moment formula from Roshan, Jones and Greenman (2014),

Theorem 3.3, taking $n = k$ and $\mu_1 = 1 - \mu_0 = 2\theta$; interestingly, a quite different approach is taken in that paper.

2.4. *Enrichment and Bayes rule.* The development so far has emphasized probabilities that condition in some way on clonotype size. Next we layer in a distribution on that size itself; the stochastic evolution of a specific clonotype σ induces a distribution on the size $N_\sigma(t_{\text{obs}})$ at observation time. For example, the linear pure-birth model leads to the Geometric$\{\exp(-\lambda_\sigma t_{\text{obs}})\}$ distribution,

$$P\{N_\sigma(t_{\text{obs}}) = n\} = e^{-\lambda_\sigma t_{\text{obs}}}\left(1 - e^{-\lambda_\sigma t_{\text{obs}}}\right)^{n-1}, \qquad n \geq 1 \tag{5}$$

where λ_σ is the birth rate (rate of cell division). Further, compounding over λ_σ gives the Yule-Simon law, with parameter $\rho > 0$,

$$P\{N_\sigma(t_{\text{obs}}) = n\} = \rho B(n, \rho + 1) = \frac{\rho\Gamma(\rho+1)\Gamma(n)}{\Gamma(n+\rho+1)} \approx \frac{\rho\Gamma(\rho+1)}{n^{\rho+1}}, \tag{6}$$

where the approximation improves with increasing n. This is approximately a power-law, or Zipf distribution, which has been found to fit many T-cell repertoires (e.g., Bolkhovskaya, Zorin and Ivanchenko, 2014; Desponds, Mora and Walczak, 2016; Koch et al., 2018; Gaimann et al., 2020; de Greef et al., 2020), with exponents ρ in the range 0.05 to 0.2. Other marginal distributions on $N_\sigma(t_{\text{obs}})$ may be induced by more complex stochastic dynamics, such those modeling competition and thymic pressure (Lythe and Molina-París, 2018).

Combining the forward, mutant-genotype model (4) with a size model $P\{N_\sigma(t_{\text{obs}}) = n\}$, we have by conditioning:

$$\begin{aligned} P\{N_\sigma(t_{\text{obs}}) = n | M_\sigma = 1\} &= \frac{P\{M_\sigma = 1 | N_\sigma(t_{\text{obs}}) = n\}\, P\{N_\sigma(t_{\text{obs}}) = n\}}{P(M_\sigma = 1)} \\ &= \frac{P\{N_\sigma(t_{\text{obs}}) = n\}}{P(M_\sigma = 1)}\left\{1 - \frac{\Gamma(n+1-2\theta)}{\Gamma(n+1)\Gamma(2-2\theta)}\right\}. \end{aligned} \tag{7}$$

This Bayesian inversion of (4) quantifies surrogate selection's enrichment effect in the pure-birth case. One setting is shown in Figure 3, which illustrates the suppression of probability on small clonotypes and inflation for larger ones. In that example, the median of the unconditional Geometric distribution is 6931 cells, while after conditioning on $M_\sigma = 1$, the median clonotype size shifts up to 8139 cells. This effect is not limited to the marginal Geometric law. Figures 4 show the result for a Logarithmic distribution (p.m.f. proportional to p^n/n) and a Yule-Simon law (6), respectively. Summarizing the findings for a single, developing clonotype, we have:

PROPOSITION 1. *Suppose that, regardless of the marginal distribution of $N_\sigma(t_{\text{obs}})$, each cell division in the developing clonotype σ increases the clonotype size by 1 and occurs on a random extant cell, that a non-mutant dividing cell produces one mutant descendant (w.p. 2θ) or no mutant descendants (w.p. $1 - 2\theta$), that descendants of a mutant dividing cell are both mutants, that there are no cell deaths, and that σ began with a single non-mutant cell. If M_σ indicates that a randomly sampled cell from σ at time t_{obs} is mutant, then the enrichment ratio $\phi_n := P\{N_\sigma(t_{\text{obs}}) = n | M_\sigma = 1\}/P\{N_\sigma(t_{\text{obs}}) = n\}$ is:*

$$\phi_n = \frac{1}{P(M_\sigma = 1)}\left\{1 - \frac{\Gamma(n+1-2\theta)}{\Gamma(n+1)\Gamma(2-2\theta)}\right\}.$$

Further, ϕ_n is strictly increasing and approaches $1/P(M_\sigma = 1) > 1$ as $n \longrightarrow \infty$.

Two immediate corollaries assure that: (1) there exists a crossover point n_{cross} with $\phi_n < 1$ when $n < n_{\text{cross}}$ and $\phi_n > 1$ when $n > n_{\text{cross}}$, and (2) the conditional distribution is stochastically larger than the marginal distribution, which is another perspective on the notion that mass is pushed towards larger clonotypes. In fact, monotonicity of ϕ_n amounts to saying that the marginal and conditional distributions satisfy the monotone likelihood ratio ordering, which is stronger than stochastic ordering of c.d.f.'s: $P\{N_\sigma(t_{\text{obs}}) \geq n | M_\sigma = 1\} \geq P\{N_\sigma(t_{\text{obs}}) \geq n\}$ (see Pfanzagl, 1964). Among other things, it also follows that the conditional distribution of $N_\sigma(t_{\text{obs}})$ given $M_\sigma = 1$ has larger expected value than the marginal distribution. Conceptually, learning that the sampled cell is mutant tells us that the clonotype is probably larger than we would have guessed otherwise.

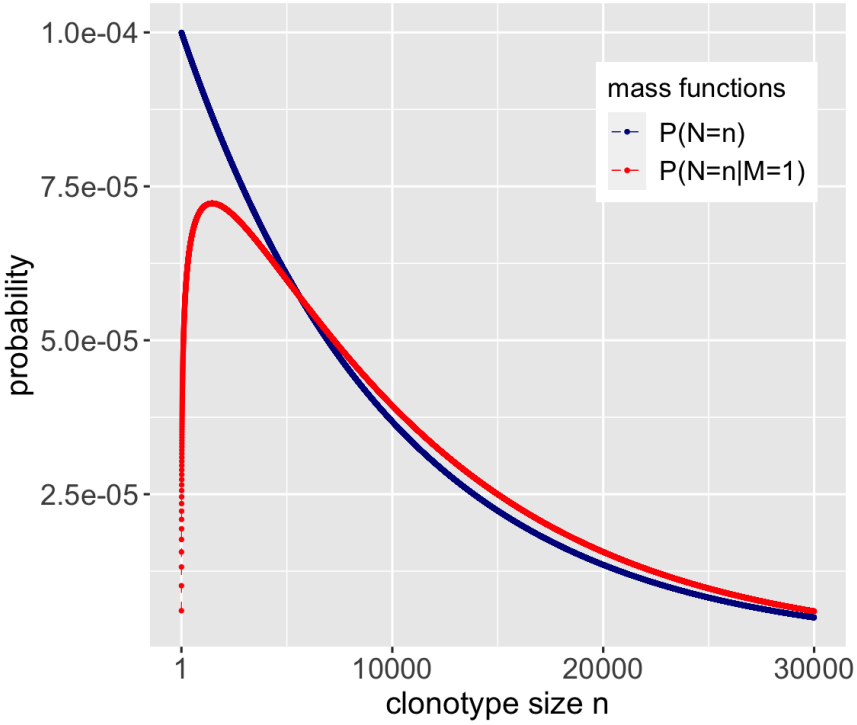

FIG 3. *$P\{N_\sigma(t_{\text{obs}}) = n | M_\sigma = 1\}$ (red) when the marginal distribution (blue) is a Geometric distribution with parameter $e^{-\lambda t_{\text{obs}}} = 10^{-4}$ and the mutation frequency $\theta = 10^{-6}$. The crossover point n_{cross} is 5624 cells.*

2.5. *Beyond pure birth.* Relaxing the no-cell-death assumption makes quantifying enrichment more difficult. Explicit calculations in one example (Appendix A) show that conditioning on $M_\sigma = 1$ does not necessarily enrich for larger clonotypes. That highly stylized example captures features of clonal expansion followed by rapid clonal decline. The intuition is that having sampled a mutant cell, we may only know that its containing clonotype is relatively old, rather than knowing this clonotype is relatively large. These two features are equivalent in the pure-birth model. To develop this intuition further, we pursue calculations in a well-behaved but general class of birth-death processes, and we find conditions within this class which assure the enrichment-for-larger-clonotypes phenomena.

At times $\tau_1 < \tau_2 < \cdots$ after τ_σ, changes $A_1, A_2, \cdots$ occur that either increase the clonotype size ($A_i = 1$) or decrease the clonotype size ($A_i = -1$), in the first case by division of a random cell, and in the latter by death of a random cell. Then at time t, the clonotype

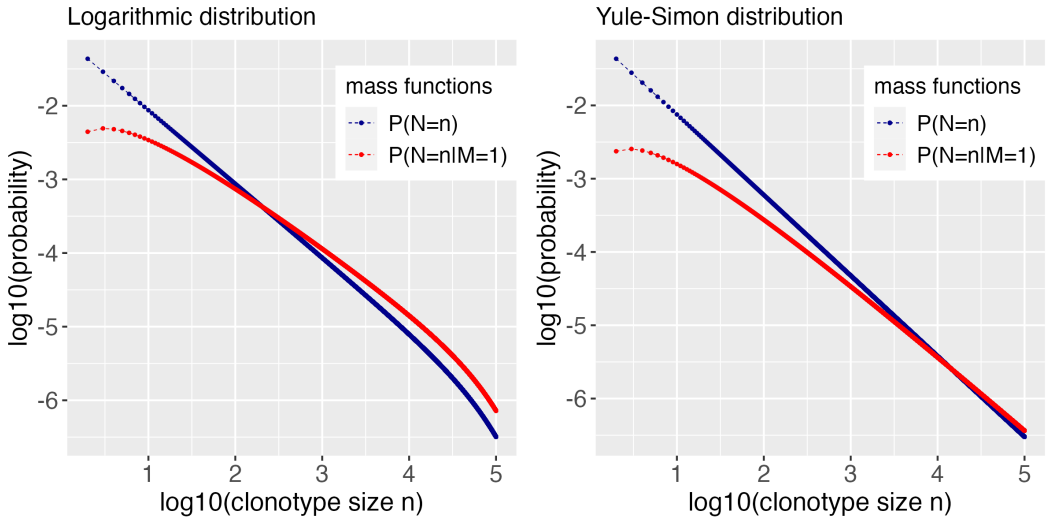

FIG 4. *$P\{N_\sigma(t_{\text{obs}}) = n | M_\sigma = 1\}$ (red) when the marginal clonotype size distribution (blue) is a Logarithmic distribution (left) or a Yule-Simon distribution (right), with parameters $p = 1 - 10^{-5}$ for Logarithmic distribution and $\rho = 0.1$ for Yule-Simon distribution. Mutation frequency $\theta = 10^{-6}$ in both cases. The crossover point n_{cross} equals to 326 cells under Logarithmic distribution, and $n_{\text{cross}} = 14270$ under Yule-Simon distribution.*

size $N_\sigma(t) = 1 + \sum_{i=1}^{I(t)} A_i$ where $\tau_{I(t)} \le t < \tau_{I(t)+1}$. We suppose this size process $N_\sigma(t)$ is not explosive, and thus only a finite number of τ_j's can occur in any finite time interval. We ask that $\{A_i\}$ be independent of event times $\tau_1 < \tau_2 < \cdots$ so that the discrete clonal history may be treated separately from questions of temporal rates of change. Further, we do not require a Markov condition, though we are mindful that having A_i conditionally independent of past changes given $\nu_{i-1} = 1 + \sum_{j=1}^{i-1} A_j$ provides for a Markovian jump chain $\nu_1, \nu_2, \cdots$, with $N_\sigma(t) = \nu_{I(t)}$ (e.g., Grimmett and Stirzaker, 2001, pg 265). Considering mutation status along the jump chain, we introduce

$$\Psi(a_1, a_2, \cdots, a_i) := P\left[M_\sigma = 1 \mid \mathcal{A}_i,\, I\left(t_{\text{obs}}\right) = i\right]$$

where $\mathcal{A}_i = \cap_{j=1}^{i} (A_j = a_j)$ tracks the specific birth-death steps; thus Ψ is the conditional mutant frequency of a cell sampled from σ just after the i birth-death steps indicated by $\mathcal{A}_i$. Obviously we cannot sample a cell from an empty clonotype, so we furthermore condition on non-extinction, i.e. $\nu_i \ge 1$ for all i. The Ψ function generalizes the pure-birth ψ_n sequence (4), which we recover with $i = (n-1)$ and all $a_j = 1$, for example.

PROPOSITION 2. *In a birth-death process as defined above, $Z_i := \Psi(A_1, A_2, \cdots, A_i)$ is non-decreasing in i. If with probability one $\sum_{j=1}^{i} 1[A_j = 1]/(j+1)$ diverges as $i \to \infty$, then Z_i converges almost surely to the limit 1, and also $E(Z_i) = P\left[M_\sigma = 1 | I(t_{\text{obs}}) = i\right]$ converges to 1. Additionally, if $\xi_{n,i} := E(Z_i | \nu_i = n)$ is non-decreasing in $i \in \{n-1, n+1, n+3, \cdots\}$ for each n, then $P\left[M_\sigma = 1 | N_\sigma(t_{\text{obs}}) = n\right] \ge \psi_n$.*

In a linear birth-death process for example, and ignoring extinction for the moment, the A_i's are i.i.d., with $P(A_i = 1) = \lambda/(\lambda+\mu)$ for birth rate $\lambda > 0$ and death rate $\mu \ge 0$. It is well known that extinction is almost sure when $\lambda \le \mu$, but also that extinction occurs with probability μ/λ as long as $\lambda > \mu$ (e.g., Grimmett and Stirzaker, 2001, pg 272). We would meet the requirements of Proposition 2 in this case; conditioning on non-extinction conditions on an event of positive probability. Note too that the divergence requirement follows immediately

from the three-series theorem (e.g., Billingsley, 1995, pg 290). We have a recursive formula for $\xi_{n,i} = E(Z_i|\nu_i = n)$; namely under the Markov condition for $\nu_1, \nu_2, \cdots$,

$$\begin{aligned}\xi_{n,i} &= P\{M_\sigma = 1 | N(t_{\text{obs}}) = n, I(t_{\text{obs}}) = i\} \\ &= w_{n,i}\xi_{n+1,i-1} + (1 - w_{n,i})\left\{\xi_{n-1,i-1}\left(1 - \frac{2\theta}{n}\right) + \frac{2\theta}{n}\right\}\end{aligned}$$

where $w_{n,i} = P(A_i = -1|\nu_i = n)$. We have not identified conditions assuring this $\xi_{n,i}$ sequence is non-decreasing in i for each n (a requirement for Proposition 2); but numerical experiments in the linear birth-death model (Figure S1) give us confidence that this condition holds in relevant settings. The final lower-bound result in Proposition 2 means that conditioning on mutant status does enrich for larger clonotypes, thus extending Proposition 1. In any case, the monotonicity of $E(Z_i)$ indicates that such conditioning enriches for older clonotypes regardless of properties of $\xi_{n,i}$.

3. Sampling from the repertoire.

3.1. *Model set up and size bias.* Calculations so far refer to the random development of a single clonotype and its internal mutation rate. More relevant to experimental data are calculations that allow for sampling from the full repertoire, and thus the simultaneous development of many clonotypes. We eschew detailed, cell-biological considerations, though we do provide necessary structural elements to allow for a distributional comparison of diversity statistics computed either from wild type or mutant T cell fractions. First we address a curious size-biased sampling effect that emerges in considering the full repertoire, in contrast to the single clonotype from Sections 2.4 and 2.5.

We focus on a single observation time t_{obs}, at which point the repertoire $\mathcal{S}$ is comprised of non-empty clonotypes $\sigma_1, \sigma_2, \cdots, \sigma_{\aleph_{\text{clo}}}$, of sizes $\mathcal{N} = \left(N_{\sigma_1}, N_{\sigma_2}, \cdots N_{\sigma_{\aleph_{clo}}}\right)$, with $\aleph_{\text{cel}} = \sum_{j=1}^{\aleph_{\text{clo}}} N_{\sigma_j}$ equal to the overall number of cells in the repertoire. We treat $\aleph_{\text{clo}}$ and $\aleph_{\text{cel}}$ as large constants, and, considering this snapshot of the repertoire, here we appreciate but do not emphasize with notation anything about the temporal, stochastic development of the clonotypes; for instance we ignore the multitude of receptors that are not extant at t_{obs}, and we therefore have $N_{\sigma_j} > 0$ for all j. We allow that some more primitive generative stochastic process may underlie the clonotype counts, but we focus on their conditional joint distribution given the total number of cells $\aleph_{\text{cel}}$ and the total number of extant clonotypes $\aleph_{\text{clo}}$, which in adult humans may be on the order of 10^{11} and 10^8, respectively. The same technical device was used by Rothman and Templeton (1980) in studying statistical properties of other assemblages, where additionally the assumption of finite exchangeability is helpful in revealing interesting system properties. We also adopt the finite exchangeability assumption for the joint mass function,

$$f_{\text{joint}}\left(n_1, n_2, \cdots, n_{\aleph_{\text{clo}}}\right) = P\left(N_{\sigma_1} = n_1, N_{\sigma_2} = n_2, \cdots, N_{\sigma_{\aleph_{\text{clo}}}} = n_{\aleph_{\text{clo}}}\right) \tag{8}$$

for counts $n_j \geq 1$, which not only simplifies the specification, but also means that joint probability masses depend on the frequency spectrum holding the *counts-of-counts*: $C(k) = \sum_\sigma 1[N_\sigma = k]$. Figure 5 realizes a small synthetic example.

To appreciate the size-bias issue, consider sampling a single cell uniformly from the repertoire, and let $S \in \mathcal{S}$ denote its clonotype identifier. We recognize that N_S, the size of the clonotype holding the sampled cell, is random owing to both the random development of the repertoire, as governed at least at the observation time by (8), and owing to the sampling of a

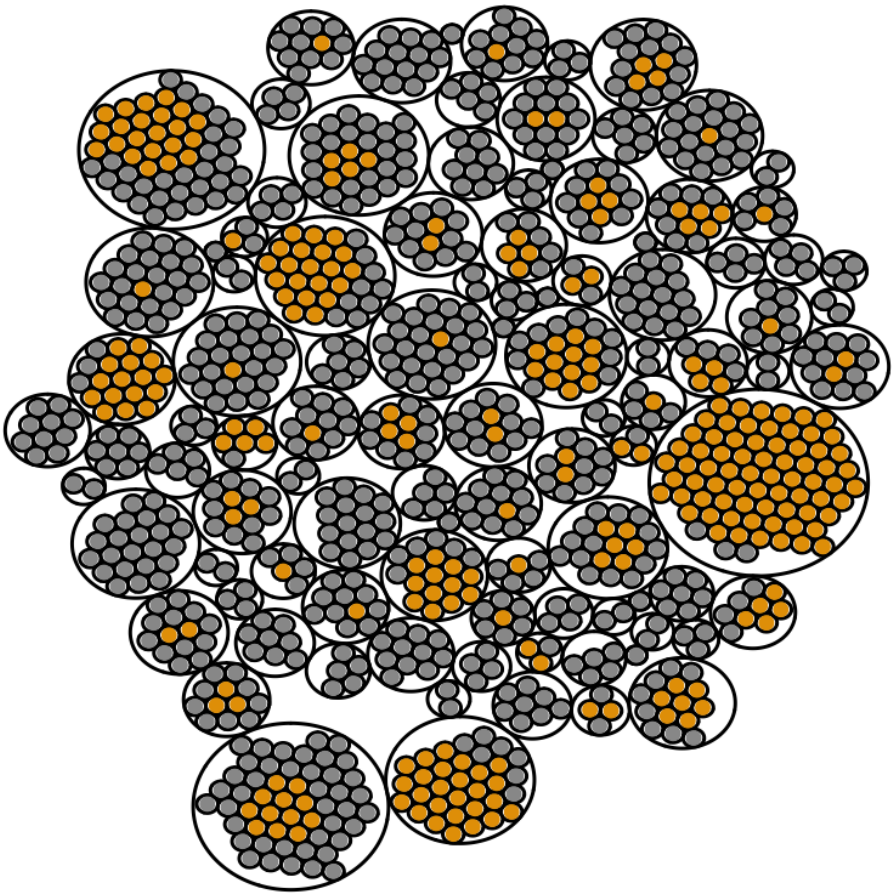

FIG 5. *Simulated repertoire of* $\aleph_{\text{cel}} = 1000$ *cells comprising* $\aleph_{\text{clo}} = 100$ *non-empty clonotypes (encasing circles). The 287 mutant cells are orange/rust, and the remaining 713 wild-type cells are grey, giving a realized mutant frequency 0.287. As predicted mathematically, the larger clonotypes have an over-representation of mutant cells. Sampling uniformly among clonotypes, the average extant clonotype size is* 10.0 *cells; given the sampled clonotype contains a mutant cell, the average clonotype size is* 16.0 *cells. On the other hand, sampling uniformly among cells, the average clonotype size of the sampled cell (i.e., with size bias) is* 23.0 *cells. The average clonotype size when sampling mutant cells, however, is even larger, at* 27.7 *cells. This synthetic data was simulated from a Bose-Einstein clone-size model and a Luria-Delbrück mutation model, with mutation frequency* $\theta = 0.05$.

cell from the repertoire. Under exchangeability, for $n \geq 1$:

$$\begin{aligned} P(N_S = n) &= \sum_{\sigma \in \mathcal{S}} P(N_S = n, S = \sigma) = \sum_{\sigma \in \mathcal{S}} P(N_\sigma = n, S = \sigma) \\ &= \sum_{\sigma \in \mathcal{S}} P(S = \sigma | N_\sigma = n)\, P(N_\sigma = n) = \sum_{\sigma \in \mathcal{S}} \left(\frac{n}{\aleph_{\text{cel}}} \right) P(N_\sigma = n) \\ &= n P(N_{\sigma_1} = n) \left(\frac{\aleph_{\text{clo}}}{\aleph_{\text{cel}}} \right). \end{aligned} \tag{9}$$

Size bias is reflected in the multiplication by n in (9). It conveys the fact that sampling a cell uniformly at random from a randomly developing repertoire is different (i.e., is biased towards larger clonotypes) than sampling a cell uniformly at random from a randomly devel-

oping clonotype. In any case, surrogate selection aims to further bias distributions towards larger clonotypes than would be obtained marginally. Before studying this enrichment, it is helpful to investigate a few exchangeable models and their relationship to well-known marginal distributions.

3.2. *Joint assemblages and limiting margins: examples.* By various compounding and conditioning operations applied to a collection of independent Poisson variates, Rothman and Templeton (1980) obtained an interesting exchangeable specification that we reconsider for (8):

$$f_{\text{joint}}(n_1, n_2, \cdots, n_{\aleph_{\text{clo}}}) \propto \prod_{j=1}^{\aleph_{\text{clo}}} \frac{\Gamma(n_j + \alpha)}{\Gamma(n_j + 1)}, \tag{10}$$

where the system-defining parameter $\alpha > 0$ reflects dynamics of the assemblage. By modifying limiting regimes for $\aleph_{\text{cel}}$, $\aleph_{\text{clo}}$, and α, Rothman and Templeton (1980), *inter alia*, recovered reference marginal distributions distinguished especially by tail behavior. For example, setting $\alpha = 1$ is the Bose-Einstein case. Sending $\aleph_{\text{clo}}/\aleph_{\text{cel}} \to \gamma_0 \in (0, 1)$ as both the numerator and denominator diverge in this case, the marginal limiting distribution of any one clonotype size is Geometric(γ_0), as in (5), which matches the pure-birth Yule tree model, with $\gamma_0 = e^{-\lambda_\sigma t_{\text{obs}}}$. Similarly, if $\alpha \to 0$, the limiting margin is the Logarithmic distribution, with p.m.f. proportional to γ_0^n/n; and if the limit of $\aleph_{\text{clo}}/\aleph_{\text{cel}}$ itself has a Beta$(\rho, 1)$ distribution, then the limiting margin is the Yule-Simon power law (6). Empirical size distributions from the Bose-Einstein simulation conform nicely to these theoretical predictions (Figure S3). These intriguing relationships provide a modeling framework allowing us to elaborate single-clonotype calculations (Section 2) into the context of full-repertoire sampling. In particular, where various conditions on the joint assemblage give rise to different limiting marginal distributions for a given clonotype's N_σ, we can similarly deduce the size-biased distribution of N_S. Details are provided in Appendix B; summarizing here, the size-biased version of the Geometric (5) has p.m.f. $n\gamma_0^2(1-\gamma_0)^{n-1}$, and the size-biased version of the Yule-Simon (6) has the p.m.f. $\rho n B(n, \rho + 2)$; see also Fig S2. We are not using these distributions for any sort of model-based inference from data; rather, we are exercising them primarily to explore implications of single versus multi-clonal analysis.

3.3. *Enrichment.* Size bias attributable to repertoire versus single-clonotype sampling does not alter the basic enrichment properties revealed in Propositions 1 and 2, except for a slight change in constants. For example, with the mutation model as in Section 2.4, and such that within each clonotype the stochastic process meets the conditions of Proposition 1, we have:

$$\frac{P(N_S = n | M_S = 1)}{P(N_S = n)} = \frac{1}{P(M_S = 1)} \left\{ 1 - \frac{\Gamma(n + 1 - 2\theta)}{\Gamma(n+1)\Gamma(2-2\theta)} \right\}$$

which is also a strictly increasing function of n that approaches limit $1/P(M_S = 1)$. The result follows from the single-clonotype sampling result (4), Bayes's rule, and the equality:

$$\begin{aligned} P(M_S = 1 | N_S = n) &= \sum_{\sigma \in \mathcal{S}} P(M_S = 1, S = \sigma \mid N_S = n) \\ &= \sum_{\sigma \in \mathcal{S}} P(M_\sigma = 1 \mid N_\sigma = n, S = \sigma)\, P(S = \sigma | N_S = n) \\ &= P(M_\sigma = 1 \mid N_\sigma = n) \qquad \text{for any } \sigma \in \mathcal{S}. \end{aligned} \tag{11}$$

By analogy, Proposition 2 may also be extended to sampling from the full repertoire. In summary,

PROPOSITION 3. *If clonotype sizes at observation time t_{obs} are exchangeable, as in (8), and if each individual clonotype evolves to its size at t_{obs} according to the dynamics in Proposition 1 or Proposition 2, then conditional on mutation $M_S = 1$ of a cell randomly drawn from the full repertoire, the enrichment ratio $P(N_S = n|M_S = 1)/P(N_S = n)$ eventually exceeds 1 for sufficiently large n.*

The enrichment phenomenon is illustrated in the synthetic repertoire in Figure 5, which shows mutant and wild-type subclones of various clonotypes, and highlights how sampling the mutant fraction would bias towards larger clonotypes.

3.4. *Mutant Frequency.* A random cell from the repertoire is more likely to be mutant than a random cell from any specific, randomly developing clonotype: $P(M_S = 1) > P(M_\sigma = 1)$, which we confirm in the Appendix C by a calculation similar to (9). This mutant frequency $P(M_S = 1)$ is of independent interest, and can be estimated by various dilution assays. As reviewed in Kaitz et al. (2022), the mutant frequency is different from the mutation frequency θ. The former considers the rate at which mutant cells are found in a sample from the repertoire; the latter is the rate that mutations emerge among cell divisions in a developing clonotype. Table S2 offers some numerical results for the Bose-Einstein assemblage.

3.5. *Diversity statistics.* An important motivation for the preceding theoretical calculations is to understand the impact of surrogate selection on statistics from a random sample from a repertoire. Suppose the amount of sampled material from one subject is a fraction $\epsilon = n_{\text{samp}}/\aleph_{\text{cel}}$ of the entire repertoire, and let X_σ record the number of cells within the sample of n_{samp} cells that have receptor σ. Conditional upon the clonotype sizes, we treat this empirical frequency as Poisson distributed, considering typical experimental settings and the relative rarity of individual clonotypes (e.g., Sepúlveda, Paulino and Carneiro, 2010). Thus,

$$X_\sigma | \mathcal{N} \sim \text{Poisson}\left\{\epsilon N_\sigma\right\}. \tag{12}$$

The number of clonotypes represented by k cells in the sample is $Y_k = \sum_\sigma 1[X_\sigma = k]$; most diversity statistics are computed from these occupancy counts, $\{Y_k\}$ (e.g., Lande, 1996; Zhang and Zhou, 2010; Chiffelle et al., 2020). The most simple one is $\mathcal{D} = \sum_{k=1}^{n_{\text{samp}}} Y_k$, which is the number of distinct clonotypes observed in the sample. Note also $n_{\text{samp}} = \sum_k kY_k$. Recognizing $\mathcal{D} = \sum_\sigma 1[X_\sigma > 0]$, it is immediate from exchangeability that:

$$E(\mathcal{D}) = \aleph_{\text{clo}} \left\{ 1 - \sum_{n \geq 1} e^{-n\epsilon} P(N_\sigma = n) \right\}, \qquad \text{for any one } \sigma. \tag{13}$$

Using characteristic functions, we may compute expected diversity directly for the reference marginals. For example, taking the limiting Geometric margin for $P(N_\sigma = n)$ noted in Section 3.2,

$$E(\mathcal{D}) = \aleph_{\text{clo}} \left\{ 1 - \frac{\gamma_0}{e^\epsilon - (1 - \gamma_0)} \right\}. \tag{14}$$

If $N_\sigma \sim \text{Log}(p)$, then,

$$E(\mathcal{D}) = \aleph_{\text{clo}} \left\{ 1 - \frac{\log(1 - pe^{-\epsilon})}{\log(1 - p)} \right\}. \tag{15}$$

For Yule-Simon marginal distribution with parameter ρ, we get,

$$E(\mathcal{D}) = \aleph_{\text{clo}} \left\{ 1 - \frac{\rho e^{-\epsilon}}{\rho + 1} \, {}_2F_1\left(1, 1; \rho + 2; e^{-\epsilon}\right) \right\} \tag{16}$$

where ${}_2F_1(a, b; c; z)$ is the Gaussian hypergeometric function. In typical repertoires, we expect parameter settings assuring high diversity, such that $E(\mathcal{D})$ is relatively close to n_{samp}.

Surrogate selection enables direct sampling from the mutant fraction, and our formalism allows a quantitative assessment of the selection effect on expected sample properties. By enriching for larger clonotypes, surrogate selection would seem to lead to fewer cells from very small clonotypes, and thus less diverse samples. Here we confirm that property. Set $\tilde{\epsilon} = n_{\text{samp}} / [\aleph_{\text{cel}} P(M_S = 1)]$, which is an amount larger than ϵ that is sufficient to produce, in expectation, n_{samp} mutant cells from the repertoire. These cells arise from the clonotypes according to sample counts $\tilde{X}_\sigma$, which, given the total numbers of mutant counts across the repertoire, $\tilde{\mathcal{N}} = \left\{\tilde{N}_\sigma\right\}$, then satisfy

$$\tilde{X}_\sigma \Big| \tilde{\mathcal{N}} \sim \text{Poisson}\left\{\tilde{\epsilon}\tilde{N}_\sigma\right\}. \tag{17}$$

The mutant sample, which in expectation has the same number of mutant cells as the total number of cells in the full-repertoire sample, has its own diversity, $\tilde{\mathcal{D}} = \sum_\sigma 1[\tilde{X}_\sigma > 0]$. By manipulating the probability generating function of the Luria-Delbrück distribution, and also leveraging results in Roshan, Jones and Greenman (2014), we find explicit formulas for the expected diversity among mutant-sampled cells.

PROPOSITION 4. *In the pure-birth, Yule tree model for clonotype development, with a* Geometric(γ_0) *distribution for each clonotype size at observation time, and with mutation frequency θ as in Proposition 1, the mutant sample has expected diversity:*

$$E(\tilde{\mathcal{D}}) = \aleph_{\text{clo}} \left\{1 - \frac{\gamma_0}{(1-e^{\tilde{\epsilon}})\{1-e^{-\tilde{\epsilon}}(1-\gamma_0)\}^{2\theta} + e^{\tilde{\epsilon}}\{1-e^{-\tilde{\epsilon}}(1-\gamma_0)\}}\right\}.$$

Alternatively, in case the clonotype-size distribution is Logarithmic(p)*, then the expected diversity is:*

$$E(\tilde{D}) = \aleph_{\text{clo}} \left\{1 - \frac{2\theta \log(1-pe^{-\tilde{\epsilon}}) - \log\left[(1-e^{\tilde{\epsilon}})(1-pe^{-\tilde{\epsilon}})^{2\theta} + e^{\tilde{\epsilon}} - p\right]}{-(1-2\theta)\log(1-p)}\right\}.$$

In either case, $E(\tilde{\mathcal{D}}) < E(\mathcal{D})$ as long as $\theta \in (0, \epsilon/2)$.

Thus in two reference models, Proposition 4 expresses the precise effect of surrogate selection on repertoire sample diversity; Figure 6 provides a numerical illustration. The result extends to more general distributions by mixing. For example, if conditional upon γ_0 the clonotype sizes are Geometric(γ_0), and if $\gamma_0 = \exp(-W)$ for $W \sim \text{Exp}(\rho)$, then marginally the clonotype size is Yule-Simon distributed with parameter ρ, and the expected diversity bound carries through the expectation: $E\left\{E\left(\mathcal{D} - \tilde{\mathcal{D}} \mid \gamma_0\right)\right\} > 0$.

3.6. *Somatic burden.* Our calculations emphasize mutation status at some special locus (like HPRT) for which experimental assays provide for ready sampling of cells within that mutant fraction of the repertoire. Yet the calculations also inform an analysis of more general mutational signatures carried by sampled T cells. Intuitively, there may be a lot of information, for example about prior antigen exposure, that is recorded in present genomic state of sampled T cells, whether or not we consider mutations for an *in vitro* selection assay.

A T cell sampled randomly from the repertoire resides in a random clonotype S of size N_S. At any genomic locus g within a host of measurable sites $\mathcal{G}$, this cell has mutation status $M_{S,g}$ relative to its prethymic state. We are thinking

$$M_{S,g} = 1\left[\text{locus } g \text{ in sampled cell has incurred a somatic mutation}\right],$$

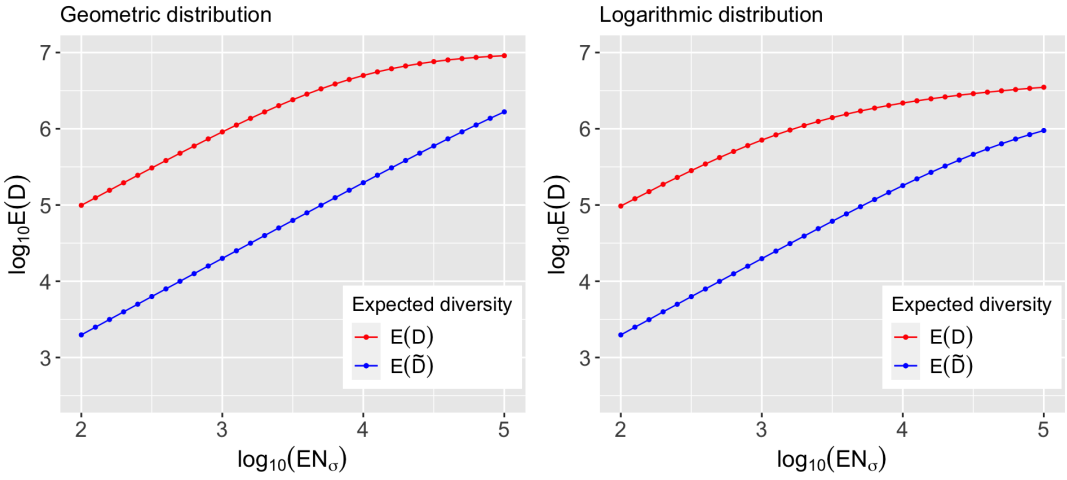

FIG 6. *Comparison of expected diversity scores between sampling from whole repertoire or just the mutant fraction, under various Geometric (left) and Logarithmic (right) distributions. The range of Geometric parameter γ_0 and logarithmic parameter p is determined to match a clonotype of approximately 10^2 to 10^5 cells, in expectation. Other parameters are fixed as sampling fraction $\epsilon = 10^{-4}$, overall number of clonotypes $\aleph_{\text{clo}} = 10^7$ and mutation probability in each division $\theta = 10^{-6}$. Expected diversity is always lower in the mutant fraction, in line with Proposition 4*

which opens us up to a genome-wide spectrum of mutations, rather than changes at a single, surrogate-selection-driving locus. To this end, we define a sampled cell's *somatic burden* L to be the summation of $M_{S,g}$ over all $g \in \mathcal{G}$. We find it convenient to consider a sequence of collections $\mathcal{G}^1, \mathcal{G}^2, \cdots$, approaching $\mathcal{G}$, with $\mathcal{G}^m$ containing m loci, and for which at step m, $P\left(M^m_{S,g} = 1 | N_S = n\right) = \psi_n(\theta^m_g)$ for locus-specific mutation frequency θ^m_g, and with ψ_n as in (3) but now highlighting its dependence on mutation frequency. This formula works in the pure-birth model structure thanks to Proposition 1 and the exchangeability in (8). Within this framework, we have the step-m burden $L^m = \sum_{g \in \mathcal{G}^m} M^m_{S,g}$.

PROPOSITION 5. *If clonotypes satisfy the regularity conditions in Proposition 1, if clonotype sizes are exchangeable as in (8), and if $\lambda^m = \sum_{g \in \mathcal{G}^m} \theta^m_g \longrightarrow \lambda$ as $m \longrightarrow \infty$ for some $\lambda > 0$, then*

$$\lim_{m \to \infty} E(L^m \mid N_S = n) = 2\lambda(H_n - 1) = \lambda \psi'_n(0) \tag{18}$$

where H_n is the n^{th} harmonic number and $\psi'_n(\theta) = d\psi_n(\theta)/d\theta$.

Put another way, the expected number of post-thymic somatic mutations in a T cell is approximately proportional to the logarithm of that cell's clonotype size, at least under the stated regularity conditions. Single-cell sequencing studies provide a means to measure L on sampled cells, and also to associate that somatic burden with clonotype size, as we investigate next.

4. Empirical studies.

4.1. *Somatic burden.* Single-cell sequencing technologies provide an exciting window into the dynamics of the T cell repertoire. Here we reanalyze publicly available data reported by 10x Genomics on samples from 7 different T cell repertoires, including 5 peripheral blood

mononuclear cell (PMBC) samples from healthy human donors, a melanoma patient and a lung cancer patient. Supplementary Material, Appendix F, summarizes the data resources and provides additional details on our analysis pipeline. In every case, the repertoire sampling and prior analysis provided both the T cell receptor (TCR) sequence and single cell whole-transcriptome RNA-seq on thousands of cells. The TCR sequence information allows us to cluster cells into clonotypes. Our interest in somatic burden puts quite different demands on the RNA-seq data than the original studies. Rather than derive transcript abundance, we repurpose the RNA-seq reads to report on underlying somatic mutations that must have emerged in the genomic DNA. Following the workflow in Edwards et al. (2022), and using the GATK pipeline for genomic-variant calling (McKenna et al., 2010; Auwera and O'Connor, 2020), we computed single-cell-expressed single-nucleotide-variant calls (sce-SNVs) from the aligned read data using `Mutect2` (Cibulskis et al., 2013; DePristo et al., 2011), applied consistently across the different repertoires. Details for SNV calling are in Appendix F, but we note here that to focus better on post-thymic somatic variants, we filtered any calls that would have appeared in more than one clonotype. In total over the 7 repertoires, we measured 30257 cells that resided in 27758 clonotypes, and which altogether presented 1609 post-thymic sce-SNVs.

Figure 7 summarizes average somatic burden as a function of clonotype size for one repertoire. Though not statistically significant, it shows an intriguing increase in estimated mean burden with increasing clonotype size, just as predicted by Proposition 5. Not all data sets show as clear a trend (Table 1), though in a meta-analysis which combines the 7 repertoires, we see stronger evidence of an increase in expected burden with clonotype size. We applied a linear model to cell-level data, with response the measured burden, and with an adjusted clonotype size predictor, where the adjustment accounts for the different sampling rates across the repertoires. We estimate $\hat{\beta} = 0.6$ SNVs per unit increase in logarithm of clonotype size. A stratified permutation, which shuffles cells between clonotypes within repertoires, gives a modest p-value of 0.02 on this clonotype-size effect. Further details are in Figure S5.

4.2. *Melanoma case studies.* We reconsider surrogate selection data presented in Zuleger et al. (2020), and we focus here (Table 2) on a metastatic melanoma patient for whom repertoire sampling was performed repeatedly over the course of what turned out to be a successful immunotherapy treatment. As the table shows, the HPRT wild-type (WT) samples have greater sample diversity than the HPRT mutant (MT) samples, which have passed *in vitro* selection.

The mass culture conditions and cDNA sequencing approach used by Zuleger et al. (2020) affect the distribution of counts in Table 2, making them over-dispersed compared to ideal cell counts. Assays based upon single-cell-derived isolates precisely count wild-type and HPRT mutant cells, rather than cDNAs, and are not subject to additional variance caused by in-vitro growth effects. However they are more labor intensive than mass cultures and provide less overall sequencing data. Table 3 summarizes such data from the peripheral blood of 11 subjects studied in Zuleger et al. (2011). In all cases the HPRT surrogate selected samples are less diverse than the wild-type cells, as predicted by the enrichment calculations in Section 3.5.

5. Concluding Remarks. Gaining a better understanding of the adaptive immune system is a central focus of contemporary biomedical research, considering that system's role in health and disease. We seek clinically useful methods to identify T cells that may be responding to antigens presented by melanoma, but it is challenging to recognize a patient's disease-specific antigens, and it is also difficult predict the antigens to which a given T cell receptor will bind. Research on both these frontiers is important and will capitalize on advances in the data sciences (e.g., Lu et al., 2021; Li et al., 2021). In any case, techniques that

TABLE 1

Somatic burden of cells by clonotype size (rows), derived from seven T cell repertoire samples (columns) made publicly available by 10x Genomics. *Details of the data resources are in Supplementary Table S3. We repurposed the single-cell RNA-seq reads to infer somatic variants and compute somatic burden counts per cell (average burden in upper table, SNVs/cell); and we used the reported TCR sequences to partition cells into clonotypes (numbers of clonotypes in bottom table).*

Clonotype size	20K	10K	SC5K	PBMC3	Controller	Melanoma	Lung
1	0.018	0.017	0.076	0.042	0.019	0.057	0.390
2	0.002	0.005	0.103	0.043	0.029	0.121	0.245
3	0	0	0	0	0	0.035	0.407
4	0	0	-	0.042	0	0.278	0.667
5	0	0	0	0	0	0	0.400
6	0	0	0	2.167	0	-	0.292
7	0	-	0	0	0	-	0.429
8	0	0	3.000	0	-	-	0.875
9	0	-	0	-	-	0	0.444
10	0	-	-	0	0	-	0.200
11	0	-	0	-	0	0	0.455
12	-	0	-	-	0	0	0.292
13	-	-	-	-	-	0	-
14	0	-	0.429	-	-	0	-
17	-	-	-	0	-	-	0.588
19	0	-	-	-	-	0	-
$[20, 40]$	0.100	0	-	-	0	-	1.283
> 40	0	-	0.170	0.171	-	-	-

Clonotype size	20K	10K	SC5K	PBMC3	Controller	Melanoma	Lung
1	8395	4211	1643	5659	4118	1097	1315
2	239	111	39	278	123	66	108
3	39	35	8	33	23	19	27
4	13	6	-	6	6	9	12
5	15	5	1	4	5	3	3
6	7	2	2	1	1	-	4
7	5	-	2	2	2	-	2
8	6	1	1	4	-	-	1
9	2	-	1	-	-	1	2
10	1	-	-	2	1	-	2
11	2	-	1	-	1	1	1
12	-	1	-	-	1	1	2
13	-	-	-	-	-	1	-
14	1	-	1	-	-	1	-
17	-	-	-	2	-	-	1
19	1	-	-	-	-	2	-
$[20, 40]$	1	1	-	-	1	-	2
> 40	1	-	1	1	-	-	-

TABLE 2

Empirical repertoire diversity in wild-type and HPRT mutant fractions, derived from sequencing TCR cDNAs from mass cultures obtained at 5 time-points on one melanoma patient

Time point	Total reads	WT unique / reads	MT unique / reads
1	108722	2840 / 58896	158 / 49826
2	111652	4587 / 53435	182 / 58217
3	98834	2709 / 49799	156 / 49035
4	87804	2091 / 52277	84 / 35527
5	98286	2209 / 51711	133 / 46575

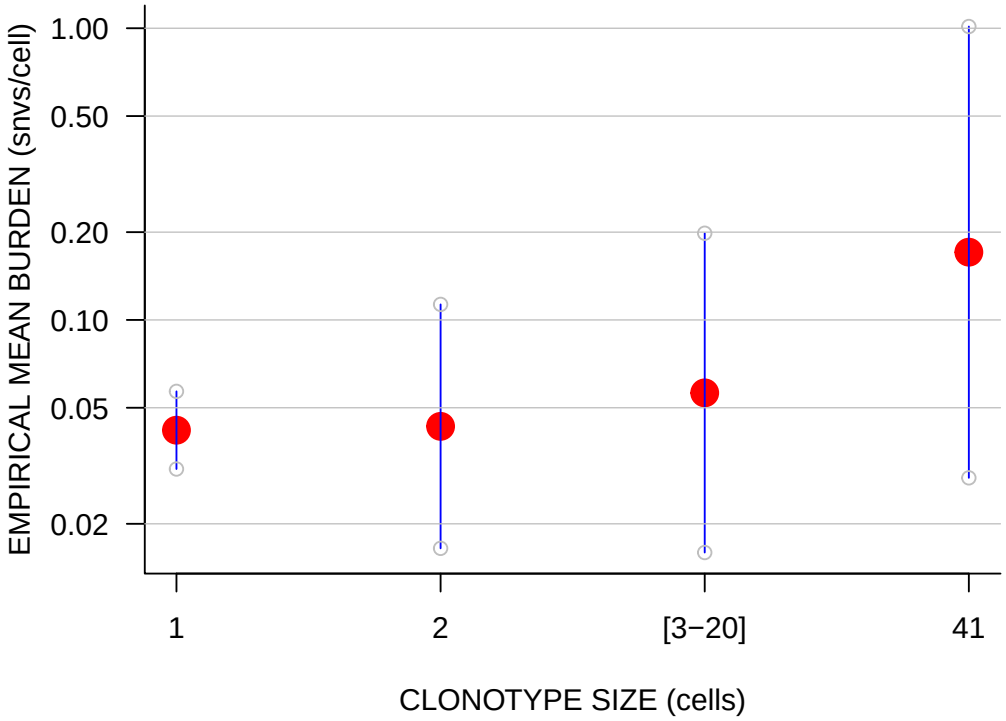

FIG 7. *Association of average somatic burden with clonotype size, in the PBMC3 repertoire. There are* 5659 *singleton clonotypes,* 278 *duplexes, and a total of* 22 *clonotypes with sizes greater than* 2*. The largest clonotype contains* 41 *cells. Clonotypes of size* 3 *to* 20 *cells are combined together as a single class considering the small sample size. Pointwise 95% confidence intervals are computed from a quasi-Poisson generalized linear model.*

TABLE 3

Empirical repertoire diversity in wild-type and HPRT mutant fractions, derived from single-cell isolate data on seven melanoma patients and four healthy donors. *Subjects 1, 2, 3, 5, 6, 9, 13 are melanoma patiens; Subjects 26, 29, 30, 32 are healthy donors. Subjects are sorted by the number sequenced T cell receptors.*

Subject	# T cells	WT unique / cells	MT unique / cells
5	122	19 / 19	102 / 103
2	114	49 / 49	61 / 65
1	101	31 / 32	45 / 69
32	95	54 / 54	30 / 41
26	81	36 / 36	44 / 45
3	79	17 / 17	55 / 62
30	69	39 / 39	29 / 30
13	69	23 / 23	43 / 46
29	56	36 / 36	19 / 20
9	50	11 / 11	23 / 39
6	26	18 / 18	8 / 8

could readily enrich a lymphocyte sample for T cells responsive to disease-relevant antigens would have a variety of practical applications. The present work provides a statistical basis to the use of surrogate selection, which aims to enrich lymphocyte samples for disease-relevant

cells by recognizing that prior clonal expansions may be associated with the accumulation of neutral somatic alterations. Relatively straightforward assays, like HPRT and PIG-A, are available to filter cells having incurred some convenient somatic alteration. Earlier studies have compared selected and unselected cell populations, using both standard and novel statistical tools to account for sources of variation affecting cell phenotypes (e.g., Pei et al., 2014; Zuleger et al., 2020). No prior studies have considered the stochastic basis of surrogate selection itself, and this problem has been the central focus of the present paper.

We treat the stochastic development of a single clonotype and demonstrate that conditioning on a mutant sampled cell enriches for larger clonotypes in a class of birth-death processes (Propositions 1 and 2). We extend the development to exchangeable collections of clonotypes (Proposition 3), accounting for the size bias and complexity of real repertoires. We study the effects of selection on the sampling distribution of a commonly computed diversity statistic (Proposition 4). Looking beyond selection, we investigate the accumulation of neutral somatic mutations across the genome, and show how the same modeling calculations demonstrate that cells in older, expanded clonotypes are expected to carry a greater mutation burden. All these theoretical predictions are accompanied by empirical results both from surrogate selection studies and recent single-cell sequencing projects. If there would be a single take-home message it would be that we have resolved the sampling phenomenon exemplified in the simulated data of Figure 5. Interestingly, cells sampled from this synthetic repertoire are associated with larger clonotypes when we condition on them being mutant, even though mutation events are completely neutral. Moreover, we hope that the quantitative characterizations developed here will provide a basis for more informed statistical analysis of T cell data sets and the planning of immunological experiments.

REFERENCES

ALBERTINI, R. J. (2001). HPRT mutations in humans: biomarkers for mechanistic studies. *Mutation Research/Reviews in Mutation Research* **489** 1-16.

ALBERTINI, R. J., CASTLE, K. L. and BORCHERDING, W. R. (1982). T-cell cloning to detect the mutant 6-thioguanine-resistant lymphocytes present in human peripheral blood. *Proceedings of the National Academy of Sciences* **79** 6617–6621.

ALBERTINI, R. J., NICKLAS, J. A., O'NEILL, J. P. and ROBISON, S. H. (1990). In vivo somatic mutations in humans: measurement and analysis. *Annual review of genetics* **24** 305–326.

ALDOUS, D. (1996). Probability Distributions on Cladograms. In *Random Discrete Structures* 1–18. Springer.

ANGERER, W. P. (2001). An explicit representation of the Luria–Delbrück distribution. *Journal of mathematical biology* **42** 145–174.

AUWERA, G. V. D. and O'CONNOR, B. D. (2020). *Genomics in the cloud: using Docker, GATK, and WDL in Terra*, 1st ed. O'Reilly Media, Sebastopol, CA.

BILLINGSLEY, P. (1995). *Probability and Measure*, 3rd ed. John Wiley & Sons.

BOLKHOVSKAYA, O. V., ZORIN, D. Y. and IVANCHENKO, M. V. (2014). Assessing T cell clonal size distribution: a non-parametric approach. *PLoS One* **9** e108658.

BROWN, G. G. and SHUBERT, B. O. (1984). On random binary trees. *Mathematics of Operations Research* **9** 43–65.

CHEEK, D. and ANTAL, T. (2018). Mutation frequencies in a birth–death branching process. *The Annals of Applied Probability* **28** 3922–3947.

CHIFFELLE, J., GENOLET, R., PEREZ, M. A., COUKOS, G., ZOETE, V. and HARARI, A. (2020). T-cell repertoire analysis and metrics of diversity and clonality. *Current Opinion in Biotechnology* **65** 284–295.

CIBULSKIS, K., LAWRENCE, M. S., CARTER, S. L., SIVACHENKO, A., JAFFE, D., SOUGNEZ, C., GABRIEL, S., MEYERSON, M., LANDER, E. S. and GETZ, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology* **31** 213–219.

CURRIE, J., CASTRO, M., LYTHE, G., PALMER, E. and MOLINA-PARÍS, C. (2012). A stochastic T cell response criterion. *Journal of The Royal Society Interface* **9** 2856–2870.

DANECEK, P., BONFIELD, J. K., LIDDLE, J., MARSHALL, J., OHAN, V., POLLARD, M. O., WHITWHAM, A., KEANE, T., MCCARTHY, S. A., DAVIES, R. M. and LI, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience* **10** giab008.

DE GREEF, P. C., OAKES, T., GERRITSEN, B., ISMAIL, M., HEATHER, J. M., HERMSEN, R., CHAIN, B. and DE BOER, R. J. (2020). The naive T-cell receptor repertoire has an extremely broad distribution of clone sizes. *Elife* **9** e49900.

DEN BRABER, I., MUGWAGWA, T., VRISEKOOP, N., WESTERA, L., MÖGLING, R., DE BOER, A. B., WILLEMS, N., SCHRIJVER, E. H., SPIERENBURG, G., GAISER, K. et al. (2012). Maintenance of peripheral naive T cells is sustained by thymus output in mice but not humans. *Immunity* **36** 288–297.

DEPRISTO, M. A., BANKS, E., POPLIN, R., GARIMELLA, K. V., MAGUIRE, J. R., HARTL, C., PHILIPPAKIS, A. A., DEL ANGEL, G., RIVAS, M. A., HANNA, M. et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* **43** 491–498.

DESPONDS, J., MORA, T. and WALCZAK, A. M. (2016). Fluctuating fitness shapes the clone-size distribution of immune repertoires. *Proceedings of the National Academy of Sciences* **113** 274–279.

DOBROVOLSKY, V. N., REVOLLO, J., PETIBONE, D. M. and HEFLICH, R. H. (2017). In vivo rat T-lymphocyte Pig-a assay: detection and expansion of cells deficient in the GPI-anchored CD48 surface marker for analysis of mutation in the endogenous Pig-a gene. In *Drug Safety Evaluation* 143–160. Springer.

DUQUE, D. F. L., MOLINA-PARIS, C., LYTHE, G., GARCIA, M. L., THOMAS, P. G. and GAEVERT, J. (2020). Stochastic modelling of the T cell repertoire with epitope affinity.

EDWARDS, N., DILLARD, C., PRASHANT, N. M., HONGYU, L., YANG, M., ULIANOVA, E. and HORVATH, A. (2022). SCExecute: custom cell barcode-stratified analyses of scRNA-seq data. *Bioinformatics* **39**. btac768.

ELHANATI, Y., SETHNA, Z., CALLAN JR, C. G., MORA, T. and WALCZAK, A. M. (2018). Predicting the spectrum of TCR repertoire sharing with a data-driven model of recombination. *Immunological reviews* **284** 167–179.

FAIRFAX, B. P., TAYLOR, C. A., WATSON, R. A., NASSIRI, I., DANIELLI, S., FANG, H., MAHÉ, E. A., COOPER, R., WOODCOCK, V., TRAILL, Z., AL-MOSSAWI, M. H., KNIGHT, J. C., KLENERMAN, P., PAYNE, M. and MIDDLETON, M. R. (2020). Peripheral CD8+ T cell characteristics associated with durable responses to immune checkpoint blockade in patients with metastatic melanoma. *Nature Medicine* **26** 193–199.

GAIMANN, M. U., NGUYEN, M., DESPONDS, J. and MAYER, A. (2020). Early life imprints the hierarchy of T cell clone sizes. *Elife* **9** e61639.

GANESAN, S. and MEHNERT, J. (2020). Biomarkers for Response to Immune Checkpoint Blockade. *Annual Review of Cancer Biology* **4** 331-351.

GRIMMETT, G. and STIRZAKER, D. (2001). *Probability and Random Processes*, 3rd ed. Oxford University Press.

HILL, B. M. (1970). Zipf's Law and Prior Distributions for the Composition of a Population. *Journal of the American Statistical Association* **65** 1220–1232.

HODGKIN, P. D., DOWLING, M. R. and DUFFY, K. R. (2014). Why the immune system takes its chances with randomness. *Nature reviews Immunology* **14** 711–711.

JOMBART, T., BALLOUX, F. and DRAY, S. (2010). Adephylo: new tools for investigating the phylogenetic signal in biological traits. *Bioinformatics* **26** 1907–1909.

JONES, I. M., GALICK, H., KATO, P., LANGLOIS, R. G., MENDELSOHN, M. L., MURPHY, G. A., PLESHANOV, P., RAMSEY, M. J., THOMAS, C. B., TUCKER, J. D. et al. (2002). Three somatic genetic biomarkers and covariates in radiation-exposed Russian cleanup workers of the Chernobyl nuclear reactor 6–13 years after exposure. *Radiation research* **158** 424–442.

KAITZ, N. A., ZULEGER, C. L., YU, P., NEWTON, M. A., ALBERTINI, R. J. and ALBERTINI, M. R. (2022). Molecular Characterization of Hypoxanthine Guanine Phosphoribosyltransferase Mutant T cells in Human Blood: The Concept of Surrogate Selection for Immunologically Relevant Cells. *Mutation Research/Reviews in Mutation Research* **789** 108414.

KENDALL, D. G. (1960). Birth-and-death processes, and the theory of carcinogenesis. *Biometrika* **47** 13–21.

KENDALL, M. G. and STUART, A. (1977). *The Advanced Theory of Statistics: Distribution theory*, 4 ed. *The Advanced Theory of Statistics*. Macmillan.

KOCH, H., STARENKI, D., COOPER, S. J., MYERS, R. M. and LI, Q. (2018). powerTCR: A model-based approach to comparative analysis of the clone size distribution of the T cell receptor repertoire. *PLoS computational biology* **14** e1006571.

LANDE, R. (1996). Statistics and partitioning of species diversity, and similarity among multiple communities. *Oikos* 5–13.

LI, G., IYER, B., PRASATH, V. B. S., NI, Y. and SALOMONIS, N. (2021). DeepImmuno: deep learning-empowered prediction and generation of immunogenic peptides for T-cell immunity. *Briefings in Bioinformatics* **22**. bbab160.

LOZANO, A. X., CHAUDHURI, A. A., NENE, A., BACCHIOCCHI, A., EARLAND, N., VESELY, M. D., USMANI, A., TURNER, B. E., STEEN, C. B., LUCA, B. A., BADRI, T., GULATI, G. S., VAHID, M. R.,

KHAMENEH, F., HARRIS, P. K., CHEN, D. Y., DHODAPKAR, K., SZNOL, M., HALABAN, R. and NEWMAN, A. M. (2022). T cell characteristics associated with toxicity to immune checkpoint blockade in patients with melanoma. *Nature Medicine* **28** 353–362.

LU, T., ZHANG, Z., ZHU, J., WANG, Y., JIANG, P., XIAO, X., BERNATCHEZ, C., HEYMACH, J. V., GIBBONS, D. L., WANG, J. et al. (2021). Deep learning-based prediction of the T cell receptor–antigen binding specificity. *Nature machine intelligence* **3** 864–875.

LYNCH, W. C. (1965). More combinatorial properties of certain trees. *The Computer Journal* **7** 299–302.

LYTHE, G. and MOLINA-PARÍS, C. (2018). Some deterministic and stochastic mathematical models of naïve T-cell homeostasis. *Immunological reviews* **285** 206–217.

MAHMOUD, H. M. (1992). *Evolution of random search trees*. *Wiley-Interscience series in discrete mathematics and optimization*. Wiley, New York.

MAHMOUD, H. M. and NEININGER, R. (2003). Distribution of distances in random binary search trees. *The Annals of Applied Probability* **13** 253–276.

MCKENNA, A., HANNA, M., BANKS, E., SIVACHENKO, A., CIBULSKIS, K., KERNYTSKY, A., GARIMELLA, K., ALTSHULER, D., GABRIEL, S., DALY, M. et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20** 1297–1303.

MOLINA-PARÍS, C. and LYTHE, G. (2021). *Mathematical, Computational and Experimental T Cell Immunology*. Springer.

NICKLAS, J. A., ALBERTINI, R. J., VACEK, P. M., ARDELL, S. K., CARTER, E. W., MCDIARMID, M. A., ENGELHARDT, S. M., GUCER, P. W. and SQUIBB, K. S. (2015). Mutagenicity monitoring following battlefield exposures: Molecular analysis of HPRT mutations in Gulf War I veterans exposed to depleted uranium. *Environmental and molecular mutagenesis* **56** 594–608.

NIKOLICH-ŽUGICH, J., SLIFKA, M. K. and MESSAOUDI, I. (2004). The many important facets of T-cell repertoire diversity. *Nature Reviews Immunology* **4** 123–132.

PARADIS, E. and SCHLIEP, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35** 526–528.

PEI, Q., ZULEGER, C. L., MACKLIN, M. D., ALBERTINI, M. R. and NEWTON, M. A. (2014). A conditional predictive p-value to compare a multinomial with an overdispersed multinomial in the analysis of T-cell populations. *Biostatistics* **15** 129–139.

PENNOCK, N. D., WHITE, J. T., CROSS, E. W., CHENEY, E. E., TAMBURINI, B. A. and KEDL, R. M. (2013). T cell responses: naïve to memory and everything in between. *Advances in Physiology Education* **37** 273-283. PMID: 24292902.

PERUZZI, B., ARATEN, D. J., NOTARO, R. and LUZZATTO, L. (2010). The use of PIG-A as a sentinel gene for the study of the somatic mutation rate and of mutagenic agents in vivo. *Mutation Research/Reviews in Mutation Research* **705** 3-10.

PFANZAGL, J. (1964). On the topological structure of some ordered families of distributions. *The Annals of Mathematical Statistics* **35** 1216–1228.

RANE, S., HOGAN, T., SEDDON, B. and YATES, A. J. (2018). Age is not just a number: Naive T cells increase their ability to persist in the circulation over time. *PLoS biology* **16** e2003949.

ROSHAN, A., JONES, P. and GREENMAN, C. (2014). Exact, time-independent estimation of clone size distributions in normal and mutated cells. *Journal of The Royal Society Interface* **11** 20140654.

ROTHMAN, E. D. and TEMPLETON, A. R. (1980). A class of models of selectively neutral alleles. *Theoretical Population Biology* **18** 135–150.

SEPÚLVEDA, N., PAULINO, C. D. and CARNEIRO, J. (2010). Estimation of T-cell repertoire diversity and clonal size distribution by Poisson abundance models. *Journal of immunological methods* **353** 124–137.

SHUM, B., LARKIN, J. and TURAJLIC, S. (2022). Predictive biomarkers for response to immune checkpoint inhibition. In *Seminars in cancer biology* **79** 4–17. Elsevier.

SMITH, C. J., VENTURI, V., QUIGLEY, M. F., TURULA, H., GOSTICK, E., LADELL, K., HILL, B. J., HIMELFARB, D., QUINN, K. M., GREENAWAY, H. Y. et al. (2020). Stochastic Expansions Maintain the Clonal Stability of CD8+ T Cell Populations Undergoing Memory Inflation Driven by Murine Cytomegalovirus. *The Journal of Immunology* **204** 112–121.

STEEL, M. and MCKENZIE, A. (2001). Properties of phylogenetic trees generated by Yule-type speciation models. *Mathematical Biosciences* **170** 91–112.

STIRK, E. R., MOLINA-PARÍS, C. and VAN DEN BERG, H. A. (2008). Stochastic niche structure and diversity maintenance in the T cell repertoire. *Journal of theoretical biology* **255** 237–249.

TAVARÉ, S. (2021). The magical Ewens sampling formula. *Bulletin of the London Mathematical Society* **53** 1563-1582.

VALPIONE, S., GALVANI, E., TWEEDY, J., MUNDRA, P. A., BANYARD, A., MIDDLEHURST, P., BARRY, J., MILLS, S., SALIH, Z., WEIGHTMAN, J., GUPTA, A., GREMEL, G., BAENKE, F., DHOMEN, N., LORIGAN, P. C. and MARAIS, R. (2020). Immune awakening revealed by peripheral T cell dynamics after one cycle of immunotherapy. *Nature Cancer* **1** 210–221.

VALPIONE, S., MUNDRA, P. A., GALVANI, E., CAMPANA, L. G., LORIGAN, P., DE ROSA, F., GUPTA, A., WEIGHTMAN, J., MILLS, S., DHOMEN, N. and MARAIS, R. (2021). The T cell receptor repertoire of tumor infiltrating T cells is predictive and prognostic for cancer survival. *Nature Communications* **12** 4098.

VAN DEN BROEK, T., BORGHANS, J. A. and VAN WIJK, F. (2018). The full spectrum of human naive T cells. *Nature Reviews Immunology* **18** 363–373.

ZHAN, Y., CARRINGTON, E. M., ZHANG, Y., HEINZEL, S. and LEW, A. M. (2017). Life and Death of Activated T Cells: How Are They Different from Naïve T Cells? *Frontiers in Immunology* **8** 1809.

ZHANG, Z. and ZHOU, J. (2010). Re-parameterization of multinomial distributions and diversity indices. *Journal of Statistical Planning and Inference* **140** 1731–1738.

ZULEGER, C. L., MACKLIN, M. D., BOSTWICK, B. L., PEI, Q., NEWTON, M. A. and ALBERTINI, M. R. (2011). In vivo 6-thioguanine-resistant T cells from melanoma patients have public TCR and share TCR beta amino acid sequences with melanoma-reactive T cells. *Journal of Immunological Methods* **365** 76–86.

ZULEGER, C. L., NEWTON, M. A., MA, X., ONG, I. M., PEI, Q. and ALBERTINI, M. R. (2020). Enrichment of melanoma-associated T cells in 6-thioguanine-resistant T cells from metastatic melanoma patients. *Melanoma research* **30** 52.

SUPPLEMENTARY MATERIAL

We provide derivations, proofs, and additional modeling elements in support of findings presented in the main manuscript, "Surrogate selection oversamples expanded T cell clonotypes", by Yu, Lian, Zuleger, Albertini, Albertini, and Newton. We also provide further details regarding data preparation and analysis from Section 4 of that work. Supplementary material is organized in seven appendices indicated below; the section in parentheses refers to the numbering within the main manuscript.

APPENDICES:

A: Enrichment in single clonotype case (Sections 2.4 and 2.5)
B: Poisson induced assemblages (Section 3.2)
C: Mutant Frequency (Section 3.4)
D: Expected Diversity (Section 3.5)
E: Burden Statistics (Section 3.6)
F: Variant calling (Section 4.1)
G: Additional Figures and Tables

APPENDIX A: ENRICHMENT IN SINGLE CLONOTYPE CASE (SECTIONS 2.4; 2.5)

Pure birth case. Proposition 1 is established by the arguments in and leading up to Section 2.4. Recall the mutant frequency among cells in a pure-birth clonotype of a given size:

$$\begin{aligned} \psi_n &= P\{M_\sigma = 1 \mid N_\sigma(t_{\text{obs}}) = n\} \\ &= 1 - \frac{\Gamma(n+1-2\theta)}{\Gamma(n+1)\Gamma(2-2\theta)} \qquad n = 1, 2, \cdots \end{aligned} \tag{19}$$

as in Eq (4).

Birth-death cases.

Counter example. There are birth-death processes for which surrogate selection fails to enrich the sample for larger clones. Suppose a clonotype σ develops by a linear pure birth process up to some time t_0, and consider some fixed threshold $K \geq 3$. Suppose also that for $t > t_0$, $N_\sigma(t) = N_\sigma(t_0)$ with probability one if $N_\sigma(t_0) < K$, and $N_\sigma(t) = 1$ with probability one if $N_\sigma(t_0) \geq K$. In other words, the birth rate drops to 0 and the death rate remains 0 after t_0 if the clone size is less than K; otherwise the death rate becomes ∞ until exactly one cell survives. This is a highly stylized model in which a clonotype of size less than K at time t_0 is essentially naïve; and then after t_0 it is more like a post-activation T cell clone where only one memory T cell remains. Here, the mutant frequency may not increase along with the clone size. For example, at $t_{\text{obs}} > t_0$,

$$\begin{aligned} P\{M_\sigma = 1 \mid N_\sigma(t_{\text{obs}}) = 1\} &\geq P\{M_\sigma = 1,\, N_\sigma(t_0) \geq K \mid N_\sigma(t_{\text{obs}}) = 1\} \\ &= \sum_{k=K}^{\infty} P\{M_\sigma = 1 \mid N_\sigma(t_0) = k\}\, P\{N_\sigma(t_0) = k \mid N_\sigma(t_{\text{obs}}) = 1\}. \end{aligned}$$

The first factor in each summand does not involve $N_\sigma(t_{\text{obs}})$ explicitly, since in this event the sampled cell yielding $M_\sigma = 1$ is the same as the single surviving cell after the spate of

cell deaths. And this factor is exactly the conditional mutant frequency ψ_k for a pure-birth process, which we recall is strictly increasing in its argument. Furthermore, by the structure of the process after t_0, and for $k \geq K$,

$$\begin{aligned} P\left\{N_\sigma(t_0) = k \mid N_\sigma(t_{\text{obs}}) = 1\right\} &= \frac{P\left\{N_\sigma(t_0) = k, N_\sigma(t_{\text{obs}}) = 1\right\}}{P\left\{N_\sigma(t_0) > K\right\} + P\left\{N_\sigma(t_0) = 1\right\}} \\ &= \frac{P\left\{N_\sigma(t_0) = k\right\}}{P\left\{N_\sigma(t_0) > K\right\} + P\left\{N_\sigma(t_0) = 1\right\}} \\ &> P\left\{N_\sigma(t_0) = k\right\}. \end{aligned}$$

Taking $K = 3$ as a simple case and combining the results above, we get,

$$\begin{aligned} P\left\{M_\sigma = 1 \mid N_\sigma(t_{\text{obs}}) = 1\right\} &\geq \sum_{k=3}^{\infty} \psi_k \, P\left\{N_\sigma(t_0) = k\right\} \\ &> \psi_3 \cdot P\left\{N_\sigma(t_0) > 3\right\} \\ &= \frac{5\theta - 2\theta^2}{3} \cdot (1-p)^2, \end{aligned}$$

which at $\theta = 0.1$ and for Geometric parameter $p = 0.1$ gives the bound 0.1296, for example. On the other hand, given $N_\sigma(t_{\text{obs}}) = 2$, the mutant frequency is:

$$P\{M_\sigma = 1 \mid N_\sigma(t_{\text{obs}}) = 2\} = \psi_2 = \theta,$$

which equals to 0.1 under that parameter setting. Therefore, $P\{M_\sigma = 1 \mid N_\sigma(t_{\text{obs}}) = 1\} > P\{M_\sigma = 1 \mid N_\sigma(t_{\text{obs}}) = 2\}$ in this toy example. Furthermore, the enrichment ratio ϕ_n could be less than 1 for any feasible clone sizes, as the largest possible clone size is K when observation time $t_{\text{obs}} > t_0$, and all $\phi_n < 1$ as long as the threshold is less than the crossover point, i.e. $K < n_{\text{cross}}$.

PROOF OF PROPOSITION 2. For the mutation frequency

$$\Psi(a_1, a_2, \cdots, a_i) = P\left[M_\sigma = 1 | A_1 = a_1, A_2 = a_2, \cdots, A_i = a_i, I(t_{\text{obs}}) = i\right],$$

we first establish the useful recursion:

$$\Psi(a_1, \cdots, a_i, a_{i+1}) = \begin{cases} \Psi(a_1, \cdots, a_i) & \text{if } a_{i+1} = -1 \\ \Psi(a_1, \cdots, a_i)\left(1 - \frac{2\theta}{\nu_i + 1}\right) + \frac{2\theta}{\nu_i + 1} & \text{if } a_{i+1} = 1 \end{cases} \tag{20}$$

where $\nu_i = 1 + \sum_{j=1}^{i} a_j$ is the clonotype size after i steps. To prove (20), it is helpful to introduce X_i recording the number of mutant cells among the ν_i cells in the clonotype just after step i: so, $0 \leq X_i \leq \nu_i - 1$, recalling that the originating cell is non-mutant in our model, and new mutants appear as at most one of the two daughter cells. Owing to the sampling of a cell at random to determine M_σ,

$$\Psi(a_1, a_2, \cdots, a_i) = \frac{1}{\nu_i} E\left(X_i | A_1 = a_1, A_2 = a_2, \cdots, A_i = a_i\right).$$

Moving to step $i+1$, the distribution of X_{i+1} given past steps and A_{i+1} depends on the whether A_{i+1} is a birth ($a_{i+1} = 1$) or a death ($a_{i+1} = -1$). In the case $a_{i+1} = 1$, $\nu_{i+1} = \nu_i + 1$, and $X_{i+1} = X_i + C_{i+1}$ where C_{i+1} is a Bernoulli trial taking value 1 if the dividing cell is mutant or if the dividing cell is non-mutant but a new mutation emerges from the division. Thus, conditional on X_i and the past sequence of birth-death steps,

$$P(C_{i+1} = 1 | X_i, A_1 = a_1, \cdots, A_i = a_i) = \frac{X_i}{\nu_i} + \left(1 - \frac{X_i}{\nu_i}\right) 2\theta.$$

Taking conditional expectations to average over X_i,

$$\begin{aligned}\Psi(a_1, a_2, \cdots, a_i, 1) &= \frac{1}{\nu_{i+1}} E(X_i + C_i | A_1 = a_1, A_2 = a_2, \cdots, A_i = a_i, A_{i+1} = 1) \\ &= \frac{1}{\nu_i + 1} \left\{ \nu_i \Psi(a_1, \cdots, a_i) + \Psi(a_1, \cdots, a_i) + 2\theta \left[1 - \Psi(a_1, \cdots, a_i) \right] \right\} \\ &= \Psi(a_1, \cdots, a_i) \left(1 - \frac{2\theta}{\nu_i + 1} \right) + \frac{2\theta}{\nu_i + 1}\end{aligned}$$

For the death event $a_{i+1} = -1$, $\nu_{i+1} = \nu_i - 1$; also X_{i+1} is X_i if the death is of a non-mutant cell and equals $X_i - 1$ if the death is of a mutant cell. Taking expectations confirms that $\Psi(a_1, \cdots, a_i, -1) = \Psi(a_1, \cdots, a_i)$, and so (20) is established. Intuitively, the removal of a random cell does not change features of the remaining cells.

By convexity of combinations in recursion (20), the random sequence $Z_i := \Psi(A_1, A_2, \cdots, A_i)$ is almost surely non-decreasing in i. In fact, $Z_{i+1} = Z_i + 1[A_{i+1} = 1](1 - Z_i)(2\theta)/(\nu_i + 1)$, and so

$$\begin{aligned}Z_i &= Z_1 + \sum_{j=1}^{i-1} (Z_{j+1} - Z_j) \\ &= Z_1 + \sum_{j=1}^{i-1} 1[A_{j+1} = 1] 2\theta (1 - Z_j)/(\nu_j + 1) \\ &\geq Z_1 + 2\theta \sum_{j=1}^{i-1} 1[A_{j+1} = 1](1 - Z_j)/(j + 2)\end{aligned}$$

because the clone size $\nu_j \leq j + 1$. Monotonicity also implies that $E(Z_{i+1}|Z_i) \geq Z_i$, which means the sequence forms a submartingale. By the martingale convergence theorem (e.g., Billingsley, 1995, pg 468), Z_i converges almost surely to some limit $Z \in [0, 1]$, approaching the limit from below, and so for all j, $1 - Z_j \geq 1 - Z$. Thus

$$\begin{aligned}Z_i &\geq Z_1 + 2\theta(1 - Z) \sum_{j=1}^{i-1} 1[A_{j+1} = 1]/(j + 2) \\ &= Z_1 + 2\theta(1 - Z) \sum_{j=2}^{i} 1[A_j = 1]/(j + 1),\end{aligned}$$

where the second line just invokes a change of variables in the sum. If the limit Z is less than 1 on some realization, then $1 - Z$ does not eliminate the subsequent sum. But we had assumed divergence of this sum with probability one, which would create an impossible lower bound for Z_i. The only option is for $Z = 1$ with probability one. In other words, the mutation frequency converges to 1 in the general birth-death process as long as there are sufficiently many births.

Convergence of $E(Z_i)$ is the monotone convergence theorem (e.g., Billingsley, 1995, pg 208).

On the conditional mutant frequency, and with $I = I(t_{\text{obs}})$ for shorthand,

$$\begin{aligned}P\left[M_\sigma = 1 | N_\sigma(t_{\text{obs}}) = n\right] &= P\left[M_\sigma = 1 | \nu_I = n\right] \\ &= \sum_{i=n-1}^{\infty} P\left[M_\sigma = 1, I = i | \nu_I = n\right]\end{aligned}$$

$$= \sum_{i=n-1}^{\infty} \xi_{n,i} P\left(I = i | \nu_I = n\right)$$

$$\geq \xi_{n,n-1} \sum_{i=n-1}^{\infty} P\left(I = i | \nu_I = n\right)$$

$$= \xi_{n,n-1} = \psi_n.$$

The inequality comes from the assumption that $\xi_{n,i}$ is non-decreasing in i for each n (e.g. see Fig S1).

We note for clarity that the increasingness of $\xi_{n,i}$ in i for each n must refer to the values i that are allowable in the birth-death formalism. For example, if $n = 3$, then i must be in $2, 4, 6, \cdots$; in general i can range in $\{n-1, n+1, n+3, \cdots\}$. □

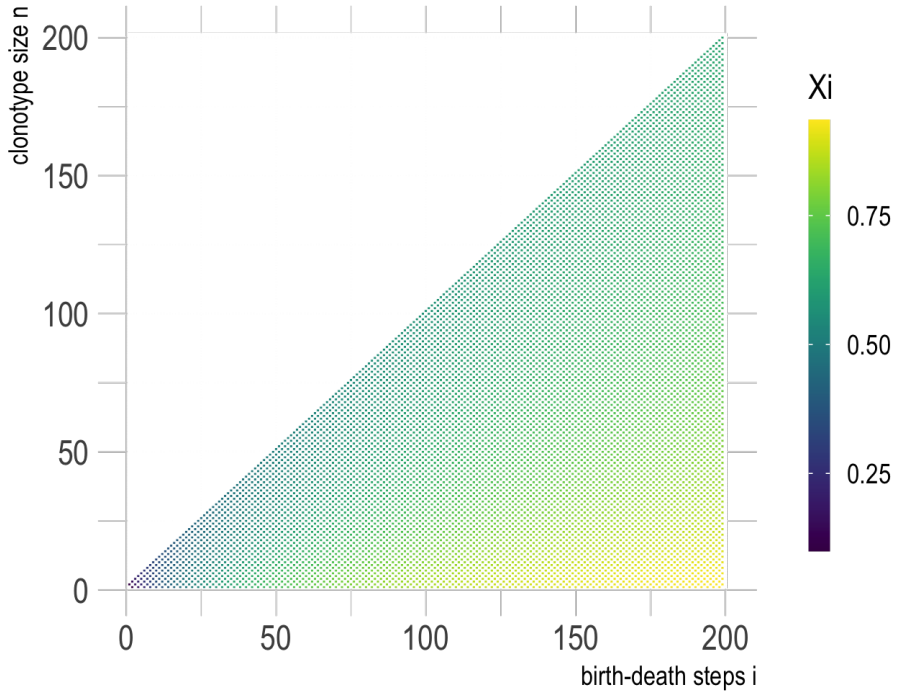

FIG S1. *$\xi_{n,i}$ for a linear birth-death process with $\lambda = 19$, $\mu = 1$, and $\theta = 0.01$. Numerically, the probabilities increase from left to right over $i \in \{n-1, n+1, n+3, \cdots\}$ for each value n, as required by Prop 2.*

APPENDIX B: POISSON-INDUCED ASSEMBLAGES (SECTION 3.2)

There are several ways in modeling the allocation of $\aleph_{\text{cel}}$ T cells among $\aleph_{\text{clo}} < \aleph_{\text{cel}}$ TCR clonotypes. The general assemblage specifications in Rothman and Templeton (1980) are informative and still relatively simple, so we develop them in the present immunological

context for completeness. We refer to this class of exchangeable models as Poisson-induced assemblages.

Unconditionally on clonotype sizes or the repertoire size, suppose the whole repertoire is generated from independent Poisson variates, with the Poisson means themselves drawn as i.i.d. from a Gamma mixture distribution. Prior to any conditioning, the resulting clonotype sizes N_k, for $k = 1, 2, \cdots, \aleph_{\text{clo}}$, are i.i.d. from a Negative Binomial distribution $\text{NB}(\alpha, p)$, with p.m.f.:

$$P(N_k = n) = \binom{n+\alpha-1}{\alpha-1}(1-p)^n p^\alpha, \quad n = 0, 1, \cdots. \tag{21}$$

Here, $p \in (0,1)$, $\alpha > 0$, and the notation follows Rothman and Templeton (1980), $\binom{a}{b} = \Gamma(a+1)/\left[\Gamma(a-b+1)\Gamma(b+1)\right]$. For valid allocations, we condition both on clonotypes being extant and on achieving a certain total repertoire size. That is, with $\aleph = (\aleph_{\text{clo}}, \aleph_{\text{cel}})$ fixed (think of them as very large constants), we will condition on the event

$$A_\aleph = \left(\bigcap_{k=1}^{\aleph_{\text{clo}}} \{N_k > 0\}\right) \cap \left(\left[\sum_{k=1}^{\aleph_{\text{clo}}} N_k\right] = \aleph_{\text{cel}}\right),$$

and we consider the induced exchangeable distribution:

$$P(N_1 = n_1, N_2 = n_2, \cdots N_{\aleph_{\text{clo}}} = n_{\aleph_{\text{clo}}} \mid A_\aleph) \tag{22}$$

The key to (22) is the probability $P(A_\aleph)$, which can be calculated through inclusion-exclusion principle:

$$\begin{aligned}
P\left(A_\aleph\right) &= \sum_{j=1}^{\aleph_{\text{clo}}} (-1)^{\aleph_{\text{clo}}-j} \binom{\aleph_{\text{clo}}}{j} P\left(\sum_{k=1}^{j} N_k = \aleph_{\text{cel}}, N_{j+1} = \cdots = N_{\aleph_{\text{clo}}} = 0\right) \\
&= \sum_{j=1}^{\aleph_{\text{clo}}} (-1)^{\aleph_{\text{clo}}-j} \binom{\aleph_{\text{clo}}}{j} \binom{\aleph_{\text{cel}}+j\alpha-1}{j\alpha-1} (1-p)^{\aleph_{\text{cel}}} p^{j\alpha} \cdot p^{(\aleph_{\text{clo}}-j)\alpha} \\
&= \sum_{j=1}^{\aleph_{\text{clo}}} (-1)^{\aleph_{\text{clo}}-j} \binom{\aleph_{\text{clo}}}{j} \binom{\aleph_{\text{cel}}+j\alpha-1}{j\alpha-1} (1-p)^{\aleph_{\text{cel}}} p^{\aleph_{\text{clo}}\alpha}
\end{aligned}$$

where the second equality is due to the sum of Negative Binomial random variables with same p also having a Negative Binomial distribution. Therefore, (22) can be calculated explicitly as:

$$\begin{aligned}
P(N_1 = n_1, N_2 = n_2, \cdots \mid A_\aleph) &= \frac{\prod_{j=1}^{\aleph_{\text{clo}}} \binom{n_j+\alpha-1}{\alpha-1}(1-p)^{n_j} p^\alpha}{\sum_{j=1}^{\aleph_{\text{clo}}} (-1)^{\aleph_{\text{clo}}-j} \binom{\aleph_{\text{clo}}}{j} \binom{\aleph_{\text{cel}}+j\alpha-1}{j\alpha-1} (1-p)^{\aleph_{\text{cel}}} p^{\aleph_{\text{clo}}\alpha}} \\
&= \frac{\prod_{j=1}^{\aleph_{\text{clo}}} \binom{n_j+\alpha-1}{\alpha-1}}{\sum_{j=1}^{\aleph_{\text{clo}}} (-1)^{\aleph_{\text{clo}}-j} \binom{\aleph_{\text{clo}}}{j} \binom{\aleph_{\text{cel}}+j\alpha-1}{j\alpha-1}}
\end{aligned} \tag{23}$$

The important special case $\alpha = 1$ is the Bose-Einstein allocation; the distribution in (23) simplifies to

$$P\left(N_1 = n_1, N_2 = n_2, \cdots \mid A_\aleph\right) = \frac{1}{\binom{\aleph_{\text{cel}}-1}{\aleph_{\text{clo}}-1}}. \tag{24}$$

This is because,

$$\sum_{j=1}^{\aleph_{\text{clo}}} (-1)^{\aleph_{\text{clo}}-j} \binom{\aleph_{\text{clo}}}{j} \binom{\aleph_{\text{cel}}+j-1}{j-1} = \binom{\aleph_{\text{cel}}-1}{\aleph_{\text{clo}}-1}$$

where both sides of the equation count how many ways there are to put $\aleph_{\text{cel}}$ indistinguishable balls into $\aleph_{\text{clo}}$ distinguishable bins, such that each bin is non-empty. The Bose-Einstein distribution assigns equal probability to all allocations of $\aleph_{\text{cel}}$ cells among $\aleph_{\text{clo}}$ non-empty clonotypes.

Introducing $C(n) = \sum_{j=1}^{\aleph_{\text{clo}}} 1[N_j = n]$ to denote the number of clonotypes comprised of n cells, the vector

$$C(1), C(2), \cdots, C(\aleph_{\text{cel}} - \aleph_{\text{clo}} + 1)$$

is called the frequency spectrum of the repertoire. These counts-of-counts are sufficient in exchangeable models since (8) depends on the clonotype sizes $\{n_k\}$ only through the frequency spectrum. In the Poisson-induced assemblages considered here,

$$P\left(N_1 = n_1, N_2 = n_2, \cdots \mid A_{\aleph}\right) \propto \prod_{n=1}^{\aleph_{\text{cel}} - \aleph_{\text{clo}} + 1} \binom{n + \alpha - 1}{\alpha - 1}^{c_n}.$$

where $c_n = \sum_j 1[n_j = n]$. By this sufficiency, the probability distribution of the frequency spectrum itself is also proportional to the above. For example, under the Bose-Einstein allocation:

$$P\left\{C(1) = c_1, C(2) = c_2, \cdots | A_{\aleph}\right\} = \frac{\aleph_{\text{clo}}!}{\prod_n c_n!} \cdot \frac{1}{\binom{\aleph_{\text{cel}} - 1}{\aleph_{\text{clo}} - 1}} \tag{25}$$

where $\sum_n c_n = \aleph_{\text{clo}}$ and $\sum n c_n = \aleph_{\text{cel}}$.

Using these facts about the joint distribution, the size N_σ of any single clonotype follows a Pólya-Eggenberger distribution, $\text{PE}(1, \aleph_{\text{clo}} - 1)$, if $\alpha = 1$:

$$\begin{aligned} P\left(N_\sigma = n \mid A_{\aleph}\right) &= E\left\{ \frac{C(n)}{\aleph_{\text{clo}}} \middle| A_{\aleph} \right\} \\ &= \frac{\binom{\aleph_{\text{cel}} - n - 1}{\aleph_{\text{clo}} - 2}}{\binom{\aleph_{\text{cel}} - 1}{\aleph_{\text{clo}} - 1}}. \end{aligned}$$

Our interest is on the limiting probability of $P\left(N_\sigma = n \mid A_{\aleph}\right)$ when $\aleph_{\text{cel}} \to \infty$ but $\aleph_{\text{clo}}/\aleph_{\text{cel}}$ converges. In the simplest scenario, the limiting ratio is a constant $0 < \gamma_0 < 1$, and so by applying Stirling's formula:

$$\lim_{\aleph_{\text{cel}} \to \infty} P\left(N_\sigma = n \mid A_{\aleph}\right) = \gamma_0 (1 - \gamma_0)^{n-1}$$

i.e. the distribution of sampled clonotype size converges to a Geometric distribution (on the support $n = 1, 2, \cdots$). Noting the original formulation (21) is Geometric when $\alpha = 1$, and on the support $n = 0, 1, 2, \cdots$ (i.e., including zero), we also see a clear connection between the Bose-Einstein specification (24) and the Geometric margin with non-zero support, and when $\aleph_{\text{clo}}$ and $\aleph_{\text{cel}}$ are large.

More generally, we may treat $\aleph_{\text{clo}}/\aleph_{\text{cel}}$ as random, and suppose it converges in distribution to Θ as $\aleph_{\text{cel}}$ diverges. For this case, Hill (1970) proved that under Bose-Einstein allocation,

$$\frac{C(n)}{\aleph_{\text{clo}}} \xrightarrow{d.} \Theta(1 - \Theta)^{n-1}$$

as $\aleph_{\text{cel}} \to \infty$. Therefore, the limiting distribution

$$\lim_{\aleph_{\text{cel}} \to \infty} P\left(N_\sigma = n \mid A_{\aleph}\right) = E\{\Theta(1 - \Theta)^{n-1}\}. \tag{26}$$

Different cases arise depending on the distribution of Θ in (26). If $\Theta \sim \mathrm{Beta}(\rho, 1)$, we arrive at the Yule-Simon limit:

$$\lim_{\aleph_{\mathrm{cel}} \to \infty} P\left(N_\sigma = n \mid A_\aleph\right) = \rho B(n, \rho + 1).$$

In other words, with Bose-Einstein allocation of $\aleph_{\mathrm{cel}}$ cells among $\aleph_{\mathrm{clo}}$ clonotypes, if the ratio $\aleph_{\mathrm{clo}}/\aleph_{\mathrm{cel}}$ converges to a Beta random variable, then the distribution of sampled clonotype size converges to a Yule-Simon distribution.

The current development allows us to directly calculate the implications of size bias caused by sampling cells from the repertoire. Suppose we sample one T cell uniformly at random from the whole repertoire, and it happens to be from random clonotype S. The size N_S of this clonotype satisfies:

$$\begin{aligned} P\left(N_S = n \mid A_\aleph\right) &= E\left\{ \frac{nC(n)}{\aleph_{\mathrm{cel}}} \middle| A_\aleph \right\} \\ &= \frac{n\binom{\aleph_{\mathrm{cel}} - n - 1}{\aleph_{\mathrm{clo}} - 2}}{\binom{\aleph_{\mathrm{cel}}}{\aleph_{\mathrm{clo}}}}. \end{aligned}$$

If $\lim_{\aleph_{\mathrm{cel}} \to \infty} \aleph_{\mathrm{clo}}/\aleph_{\mathrm{cel}} = \gamma_0$ for a positive constant γ_0, then

$$\lim_{\aleph_{\mathrm{cel}} \to \infty} P\left(N_S = n \mid A_\aleph\right) = n\gamma_0^2(1 - \gamma_0)^{n-1}, \qquad n = 1, 2, \cdots. \tag{27}$$

i.e., the limiting distribution of size-biased clonotype size is no longer Geometric, but instead is a Negative Binomial distribution, shifted by 1, that is: $1 + \mathrm{NB}(2, \gamma_0)$, with support on positive integers. Similarly, if $\aleph_{\mathrm{cel}}^{-1}\aleph_{\mathrm{clo}}$ converges in distribution to $\Theta \sim \mathrm{Beta}(\rho, 1)$ distribution, we find,

$$\begin{aligned} \lim_{\aleph_{\mathrm{cel}} \to \infty} P\left(N_S = n \mid A_\aleph\right) &= E\{\Theta^2(1 - \Theta)^{n-1}\} \\ &= \rho n B(n, \rho + 2). \end{aligned} \tag{28}$$

This limit is also a power-law distribution, with

$$\rho n B(n, \rho + 2) \approx \rho \Gamma(\rho + 2) \frac{1}{n^{\rho+1}}$$

for large n. Notice that Yule-Simon distribution p.m.f. $\rho B(n, \rho + 1)$ is approximately $\rho\Gamma(\rho + 1)n^{-\rho-1}$ for large n; the size bias effect curiously does not change the tail weight in this case. Figure S2 shows the size-bias effect on Yule-Simon distribution.

Our last comment on Bose-Einstein allocation is its equivalency to Dirichlet-Multinomial distribution $\mathrm{DirMult}(n, a)$, where $a = (a_1, \cdots, a_{\aleph_{\mathrm{clo}}})$ are the (positive) shape parameters of Dirichlet distribution. The DirMult distribution is applied in simulation of Fig 5 (main manuscript). The Bose-Einstein allocation and Dirichlet-Multinomial distribution can both be realized by Pólya-Eggenberger's urn model: suppose an urn contains $\aleph_{\mathrm{clo}}$ balls of distinct colors, at each time a random ball is drawn form the urn and replaced with two balls of the same color. Repeating the sampling and replacement procedure $\aleph_{\mathrm{cel}} - \aleph_{\mathrm{clo}}$ times will lead to Bose-Einstein allocation with $\aleph_{\mathrm{cel}}$ balls in $\aleph_{\mathrm{clo}}$ colors. In the mean time, this Pólya urn model also realizes the $\mathrm{DirMult}(\aleph_{\mathrm{cel}} - \aleph_{\mathrm{clo}}, \mathbb{1})$ distribution with $\aleph_{\mathrm{cel}} - \aleph_{\mathrm{clo}}$ trials and all shapes a_i equal to 1. Therefore, the Dirichlet-Multinomial simulation of repertoire in Fig 5 is the same as Bose-Einstein allocation of 1000 cells among 100 clonotypes, which is also the mixture of 100 marginally Geometric-sized clonotypes with 1000 cells in total, as noted previously.

The limiting cases considered above do not handle the case of the Logarithmic marginal. However, we note that taking a limit $\alpha \to 0$ in (21), and conditioning on $N_k \geq 1$, then the NB distribution converges to the Logarithmic distribution, with p.m.f. proportional to p^n/n, as in (Kendall and Stuart, 1977, page 139). Further, the joint distribution of the frequency spectrum becomes the Ewens sampling formula (Tavaré, 2021).

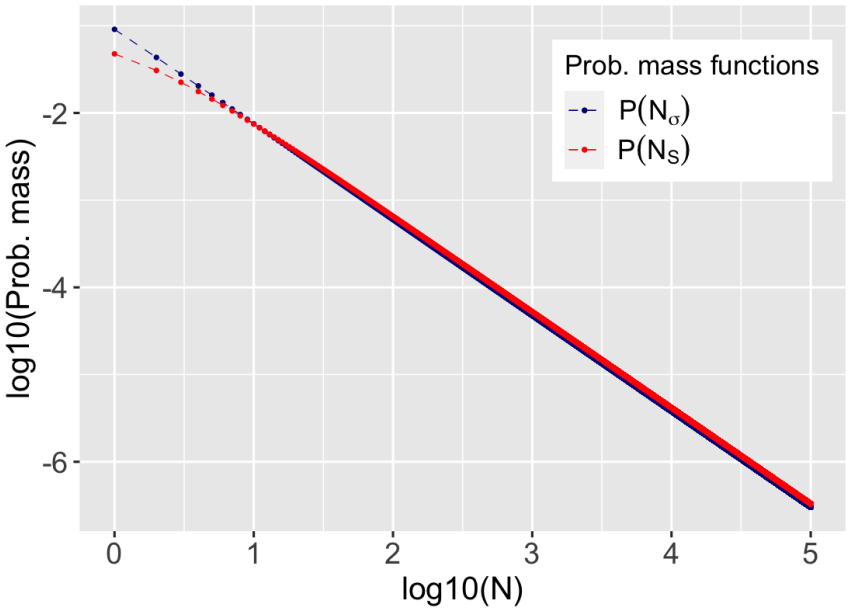

FIG S2. *Comparison of limiting distributions of clonotype sizes between size-biased sampling and uniform sampling. Here uniform sampling means uniformly sample a clonotype from the repertoire, while size-biased sampling means uniformly sample a cell from the repertoire.*

APPENDIX C: MUTANT FREQUENCY (SECTION 3.4)

First we show that size bias inflates $P(M_S = 1)$ over $P(M_\sigma = 1)$ for any clonotype σ. We work with a finitely exchangeable joint assemblage, such as in Appendix B. Recall that $\psi_n = P(M_\sigma = 1 \mid N_\sigma = n)$ is the mutant frequency defined by Luria-Delbrück distribution, as in (19). Expanding the marginal probability and using exchangeability,

$$\begin{aligned} P\left(M_S = 1 \mid A_{\aleph}\right) &= \sum_n P\left(M_S = 1 \mid N_S = n, A_{\aleph}\right) P\left(N_S = n \mid A_{\aleph}\right) \\ &= \sum_n \psi_n E\left\{ \frac{nC(n)}{\aleph_{\text{cel}}} \middle| A_{\aleph} \right\} \\ &= \sum_n n\psi_n \frac{\aleph_{\text{clo}}}{\aleph_{\text{cel}}} E\left\{ \frac{C(n)}{\aleph_{\text{clo}}} \middle| A_{\aleph} \right\}. \end{aligned}$$

The inflation of mutant frequency is reflected in the term $n\psi_n$ in the last equation, as surrogate selection will further enrich the mutants in larger clonotypes. In the Bose-Einstein case, for example,

$$\begin{aligned} P(M_S = 1 \mid \aleph_{\text{cel}}, \aleph_{\text{clo}}) &= \sum_n n\psi_n \frac{\aleph_{\text{clo}}}{\aleph_{\text{cel}}} E\left\{ \frac{C(n)}{\aleph_{\text{clo}}} \middle| \aleph_{\text{cel}}, \aleph_{\text{clo}} \right\} \\ &= \sum_n n \left(1 - \frac{\Gamma(n+1-2\theta)}{\Gamma(n+1)\Gamma(2-2\theta)} \right) \frac{\aleph_{\text{clo}}}{\aleph_{\text{cel}}} \frac{\binom{\aleph_{\text{cel}}-n-1}{\aleph_{\text{clo}}-2}}{\binom{\aleph_{\text{cel}}-1}{\aleph_{\text{clo}}-1}} \\ &= 1 - \sum_n \frac{\Gamma(n+1-2\theta)\Gamma(\aleph_{\text{cel}}-n)\Gamma(\aleph_{\text{clo}}+1)\Gamma(\aleph_{\text{cel}}-\aleph_{\text{clo}}+1)}{\Gamma(n)\Gamma(2-2\theta)\Gamma(\aleph_{\text{cel}}+1)\Gamma(\aleph_{\text{clo}}-1)\Gamma(\aleph_{\text{cel}}-\aleph_{\text{clo}}-n+2)} \end{aligned} \tag{29}$$

where the second equation is from the Bose-Einstein allocation of $\aleph_{\text{cel}}$ cells among $\aleph_{\text{clo}}$ clonotypes. The summation in (29) is finite from 1 to $\aleph_{\text{cel}} - \aleph_{\text{clo}} + 1$, and hence we can directly calculate the mutant frequency of the repertoire condition on $\aleph_{\text{cel}}$ and $\aleph_{\text{clo}}$. In reality, $\aleph_{\text{cel}}$ ranges from 10^9 to 10^{10} in the blood, and $\aleph_{\text{clo}}$ ranges from 10^6 to 10^8. Table S2 compares the marginal mutant frequency for various mutation probability θ and number of clonotypes $\aleph_{\text{clo}}$, with total number of T cells fixed at $\aleph_{\text{cel}} = 10^9$. The calculated mutant frequency will decrease as θ decreases or average clonotype size $\aleph_{\text{cel}}/\aleph_{\text{clo}}$ decreases.

APPENDIX D: EXPECTED DIVERSITY FOR LOGARITHMIC AND YULE-SIMON (SECTION 3.5)

The expected diversity can be derived directly from the characteristic function of marginal distributions, noting (13). For Logarithmic marginal distribution, $N_\sigma \sim \text{Log}(p)$, we find:

$$E(\mathcal{D}) = \aleph_{\text{clo}} \left\{ 1 - \frac{\log(1 - pe^{-\epsilon})}{\log(1-p)} \right\}.$$

For Yule-Simon marginal distribution with parameter ρ, the expected diversity is more complicated. We calculate:

$$E(\mathcal{D}) = \aleph_{\text{clo}} \left\{ 1 - \frac{\rho e^{-\epsilon}}{\rho + 1} \, {}_2F_1\left(1, 1; \rho + 2; e^{-\epsilon}\right) \right\}$$

where ${}_2F_1(a, b; c; z)$ is the Gaussian hypergeometric function.

PROOF OF PROPOSITION 4. From definition of diversity statistic $\mathcal{D}$,

$$E(\mathcal{D}) = \aleph_{\text{clo}} \left\{ 1 - E(e^{-\epsilon N_\sigma}) \right\}$$

When marginally $P(N_\sigma = n) = (1-\gamma_0)^{n-1}\gamma_0$, we have

$$E(e^{-\epsilon N_\sigma}) = \frac{\gamma_0}{e^\epsilon - (1-\gamma_0)}$$

from the characteristic function of the Geometric distribution, and hence

$$E(\mathcal{D}) = \aleph_{\text{clo}} \left\{ 1 - \frac{\gamma_0}{e^\epsilon - (1-\gamma_0)} \right\}.$$

Similarly, the key to $E(\tilde{\mathcal{D}})$ is the expectation $E(e^{-\tilde{\epsilon}\tilde{N}_\sigma})$. We refer to Theorem 3.2 in Roshan, Jones and Greenman (2014) to derive this expectation by conditioning on N_σ. That paper derives an explicit formula for the generating function of (N_σ, W_σ), where N_σ is the size of the entire clonotype and W_σ is the number of wild type cells. These sizes are defined for extant clonotypes, i.e. $N_\sigma \geq 1$. The generating function used is:

$$G(x, y) = \sum_{k=1}^{\infty} \sum_{n=1}^{\infty} P(W_\sigma = n \mid N_\sigma = k) x^{k-1} y^{n-1} \tag{30}$$

If we let $x = te^{-\tilde{\epsilon}}$ and $y = e^{\tilde{\epsilon}}$ in Eq. (30), we have:

$$\begin{aligned} H(t) = G(te^{-\tilde{\epsilon}}, e^{\tilde{\epsilon}}) &= \sum_{k=1}^{\infty} \sum_{n=1}^{\infty} P(W_\sigma = n \mid N_\sigma = k) t^{k-1} e^{-\tilde{\epsilon}(k-n)} \\ &= \sum_{k=1}^{\infty} E\left(e^{-\tilde{\epsilon}(N_\sigma - W_\sigma)} \mid N_\sigma = k\right) t^{k-1} \\ &= \sum_{k=1}^{\infty} E\left(e^{-\tilde{\epsilon}\tilde{N}_\sigma} \mid N_\sigma = k\right) t^{k-1}. \end{aligned}$$

The last equation is because of equality $\tilde{N}_\sigma = N_\sigma - W_\sigma$. If $N_\sigma \sim \text{Geom}(\gamma_0)$, we can let $t = 1 - \gamma_0$ for $H(t)$ to get the explicit formula for $E\{\exp(-\tilde{\epsilon}\tilde{N}_\sigma)\}$:

$$\begin{aligned}\gamma_0 H(1-\gamma_0) &= \sum_{k=1}^{\infty} \gamma_0 (1-\gamma_0)^{k-1} E\left(e^{-\tilde{\epsilon}\tilde{N}_\sigma} \mid N_\sigma = k\right) \\ &= \sum_{k=0}^{\infty} E\left(e^{-\tilde{\epsilon}\tilde{N}_\sigma} \mid N_\sigma = k\right) P(N_\sigma = k) \\ &= E\{e^{-\tilde{\epsilon}\tilde{N}_\sigma}\}\end{aligned}$$

From Theorem 3.2 in Roshan, Jones and Greenman (2014), there is an explicit formula under Luria-Delbrück distribution:

$$G(x,y) = (1-x)^{-2\theta}\left[1 - y\{1-(1-x)^{1-2\theta}\}\right]^{-1},$$

which leads to the formula of $H(t)$:

$$H(t) = (1-te^{-\tilde{\epsilon}})^{-2\theta}\left[1 - e^{\tilde{\epsilon}}\{1-(1-te^{-\tilde{\epsilon}})^{1-2\theta}\}\right]^{-1}$$

and hence we can get $E(e^{-\tilde{\epsilon}\tilde{N}_\sigma})$:

$$E\left(e^{-\tilde{\epsilon}\tilde{N}_\sigma}\right) = \frac{\gamma_0}{(1-e^{\tilde{\epsilon}})\left\{1-e^{-\tilde{\epsilon}}(1-\gamma_0)\right\}^{2\theta} + e^{\tilde{\epsilon}}\{1-e^{-\tilde{\epsilon}}(1-\gamma_0)\}}.$$

For the second part of Proposition 4, the diversity calculation depends on a comparison of drop-out probabilities. Define

$$\delta := P(\tilde{X}_\sigma = 0) - P(X_\sigma = 0),$$

which we see from the definition of our simple diversity statistic satisfies $E(\mathcal{D}) - E(\tilde{\mathcal{D}}) = \aleph_{\text{clo}}\phi$, and so it suffices to show that $\delta > 0$. From Poisson sampling,

$$\delta = E\left(e^{-\tilde{\epsilon}\tilde{N}_\sigma}\right) - E\left(e^{-\epsilon N_\sigma}\right),$$

To prove $\delta > 0$ for $N_\sigma \sim \text{Geom}(\gamma_0)$, we need to show that:

$$e^{\epsilon} - e^{\tilde{\epsilon}} + (e^{\tilde{\epsilon}} - 1)\{1-e^{-\tilde{\epsilon}}(1-\gamma_0)\}^{2\theta} > 0 \tag{31}$$

Since $\gamma_0 \in (0,1)$ and the left-hand-side of (31) is increasing as γ_0 increases, Eq. (31) is equivalent to:

$$e^{\epsilon} - e^{\tilde{\epsilon}} + (e^{\tilde{\epsilon}} - 1)\left(1-e^{-\tilde{\epsilon}}\right)^{2\theta} > 0. \tag{32}$$

Since function $f(x) = -x + (x-1)(1-1/x)^{2\theta}$ is a decreasing function on $x \in (1,e)$ and $\tilde{\epsilon} \in (0,1)$, it can be shown that:

$$\begin{aligned}e^{\epsilon} - e^{\tilde{\epsilon}} + \left(e^{\tilde{\epsilon}} - 1\right)\left(1-e^{-\tilde{\epsilon}}\right)^{2\theta} &\geq e^{\epsilon} - e + (e-1)\left(1-e^{-1}\right)^{2\theta} \\ &\geq e^{\epsilon} - e + (e-1)(1-e^{-1})^{\epsilon} \\ &> 0.\end{aligned}$$

where the second inequality is from the condition $\theta < \epsilon/2$ and the last inequality is due to the infimum is achieved at $\epsilon = 0$. Therefore, we have proved that $\delta > 0$ and hence $E(\tilde{\mathcal{D}}) < E(\mathcal{D})$ when $N_\sigma \sim \text{Geom}(\gamma_0)$.

In another case, if N_σ follows a Logarithmic distribution instead, with p.m.f.

$$P(N_\sigma = k) = \frac{-1}{\log(1-p)}\frac{p^k}{k}$$

for parameter $p \in (0,1)$, then the desired expectation $E(e^{-\tilde{\epsilon}\tilde{N}_\sigma})$ is:

$$\begin{aligned} E\{e^{-\tilde{\epsilon}\tilde{N}_\sigma}\} &= \sum_{k=1}^{\infty} E\left(e^{-\tilde{\epsilon}\tilde{N}_\sigma} \mid N_\sigma = k\right) P(N_\sigma = k) \\ &= \frac{-1}{\log(1-p)} \sum_{k=1}^{\infty}\sum_{n=1}^{\infty} P(W_\sigma = n \mid N_\sigma = k)\frac{p^k}{k} e^{-(k-n)\tilde{\epsilon}} \\ &= \frac{-1}{\log(1-p)} \int_0^p H(t)\mathrm{d}t \end{aligned}$$

where $H(t) = G(te^{-\tilde{\epsilon}}, e^{\tilde{\epsilon}})$ is previously defined. The integral in last equality is valid since the singularity of $H(t)$ is at:

$$t^* = e^{\tilde{\epsilon}}\{1 - (1-e^{-\tilde{\epsilon}})^{\frac{1}{1-2\theta}}\}$$

which satisfies $t^* > 1 > p$ for $0 < \theta < \epsilon/2$ and $\epsilon < \tilde{\epsilon} < 1$. Therefore, the expectation under Logarithmic distribution is:

$$E(e^{-\tilde{\epsilon}\tilde{N}_\sigma}) = \frac{2\theta\log(1-pe^{-\tilde{\epsilon}}) - \log\left[(1-e^{\tilde{\epsilon}})(1-pe^{-\tilde{\epsilon}})^{2\theta} + e^{\tilde{\epsilon}} - p\right]}{-(1-2\theta)\log(1-p)}$$

We now show that inequality $E(\mathcal{D}) > E(\tilde{\mathcal{D}})$ also holds for logarithmic distribution, i.e.,

$$\delta = \frac{2\theta\log(1-pe^{-\tilde{\epsilon}}) - \log\left[(1-e^{\tilde{\epsilon}})(1-pe^{-\tilde{\epsilon}})^{2\theta} + e^{\tilde{\epsilon}} - p\right]}{-(1-2\theta)\log(1-p)} - \frac{\log(1-pe^{-\epsilon})}{\log(1-p)} > 0$$

It is equivalent to:

$$(1-2\theta)\log(1-pe^{-\epsilon}) + 2\theta\log(1-pe^{-\tilde{\epsilon}}) > \log\left[(1-e^{\tilde{\epsilon}})(1-pe^{-\tilde{\epsilon}})^{2\theta} + e^{\tilde{\epsilon}} - p\right] \tag{33}$$

Since $\epsilon < \tilde{\epsilon} < 1$, $\log(1-pe^{-\tilde{\epsilon}}) > \log(1-pe^{-\epsilon})$, and hence (33) is true if we prove:

$$(e^{\tilde{\epsilon}} - 1)\left\{1 - (1-pe^{-\tilde{\epsilon}})^{2\theta}\right\} - p(1-e^{-\epsilon}) \le 0$$

Notice that function $f(x) = (x-1)\{1-(1-q/x)^{2\theta}\}$ is an increasing function for $1 < x < e$, therefore above inequality is equivalent to:

$$g(p) = (e-1)\left\{1 - (1-pe^{-1})^{2\theta}\right\} - p(1-e^{-\epsilon}) \le 0$$

With condition $0 < \theta < \epsilon/2$, the second order partial derivative $\frac{\mathrm{d}^2}{\mathrm{d}p^2}g(p) > 0$ for any $0 < p < 1$, therefore the supremum of $g(p)$ is achieved at $p = 0$ or 1. Since $g(0) = 0$ and $g(1) \le 0$, it is guaranteed that $g(p) \le 0$ for $0 < p < 1$. Therefore, our desired inequality (33) is true, and hence $E(\tilde{\mathcal{D}}) < E(\mathcal{D})$. □

APPENDIX E: BURDEN STATISTICS (SECTION 3.6)

PROOF OF PROPOSITION 5. With $L^m = \sum_{g\in\mathcal{G}^m} M^m_{g,\sigma}$ (and within one clonotype σ, since we are fixing clonotype size at n), we know that $E(L^m|N_\sigma = n) = \sum_{g\in\mathcal{G}^m} \psi_n(\theta^m_g)$, by the findings in Section 2, where θ^m_g is the mutation frequency associated with locus g, among the multitude in $\mathcal{G}^m$. These θ^m_g's must be fairly small, considering the convergence of $\lambda^m = \sum_{g\in\mathcal{G}^m} \theta^m_g$ to $\lambda > 0$. Thus by a Taylor expansion around $\theta = 0$,

$$\begin{aligned} E(L^m|N_\sigma = n) &= \sum_{g\in\mathcal{G}^m} \psi_n\left(\theta^m_g\right) \\ &= \sum_{g\in\mathcal{G}^m} \left\{\psi_n(0) + \theta^m_g\psi'_n(0) + o(\theta^m_g)\right\} \\ &= \psi'_n(0) \sum_{g\in\mathcal{G}^m} \left[\theta^m_g + o\left(\theta^m_g\right)\right] \\ &= 2\lambda(H_n - 1) + o(1) \end{aligned}$$

noting $\psi'_n(0) = 2(H_n - 1)$ is the derivative of $\psi_n(\theta)$ at $\theta = 0$. □

APPENDIX F: SOMATIC VARIANTS IN T CELLS FROM SCRNA-SEQ DATA

Our numerical experiments used seven publicly available 10x Genomics data sets obtained from `https://www.10xgenomics.com/resources/datasets`; the basic characteristics of these data are in Table S3. The 10x Genomics workflow generates scRNA and TCR sequencing reads by capturing and lysing single cells in Gel Beads-in-emulsion which contain a unique oligonucleotide barcode in each bead. According to the reported protocol, mRNA from lysed cells was reverse transcribed to barcoded cDNA. The cDNA was then PCR amplified to construct 5' Gene Expression library and TCR library. The TCR library consists of V(D)J segments amplified using TCR region specific primers. Sequences were then obtained from the two libraries by Illumina sequencing. Further, the clonotypes of T cells were identified by 10x Genomics by the expressed TRA and TRB genes from TCR-seq. The clonotype information of the T cells from the 'VDJ TCR - all contig annotations' CSV files in the online database were imported for our calculations.

For variant calling, we re-processed the single-cell RNA-seq data. The genome-aligned gene expression BAM files downloaded from the 10x Genomics database are used to call somatic variants by Mutect2 with a workflow adapted from Edwards et al. (2022). The BAM files were first processed by the GATK (v.4.2.6.1) module AddOrReplaceReadGroups (McKenna et al., 2010; Auwera and O'Connor, 2020). The module SplitNCigarReads then split the aligned RNA reads which include N elements because of the RNA splicing events. The reference sequences input in GATK modules are the same as versions used for alignment by the Cell Ranger pipeline, available on 10x Genomics and the Ensemble database. After that, somatic variants were called by Mutect2 (Cibulskis et al., 2013; DePristo et al., 2011) using the processed sequencing reads from all cells. We ran Mutect2 under the tumor-only mode. The public Panel of Normal was input to correct for technical artifacts. A population germline resource was also provided for Mutect2 so that the same alleles as in germline resource were not called somatic variants. The variants in the output VCF files were filtered to only keep the confident single nucleotide variants using BCFtools (v.1.15.1) from SAMtools (Danecek et al., 2021) with parameters TYPE="snp", INFO/DP (read depth) > 20 and MMQ (median mapping quality by allele) > 50. The filtered variants called from all RNA-seq data were then assigned to each single cell with the program vartrix (v.1.1.22) by re-aligning single-cell barcoded reads to each variant locus. Two matrices, each of numbers of reference

genotype reads and alternative genotype reads corresponding to cell barcodes, are generated by setting the “scoring-method" to “coverage" mode. These data were post-processed in R to associate somatic burden with clonotype sizes.

APPENDIX G: ADDITIONAL FIGURES AND TABLES

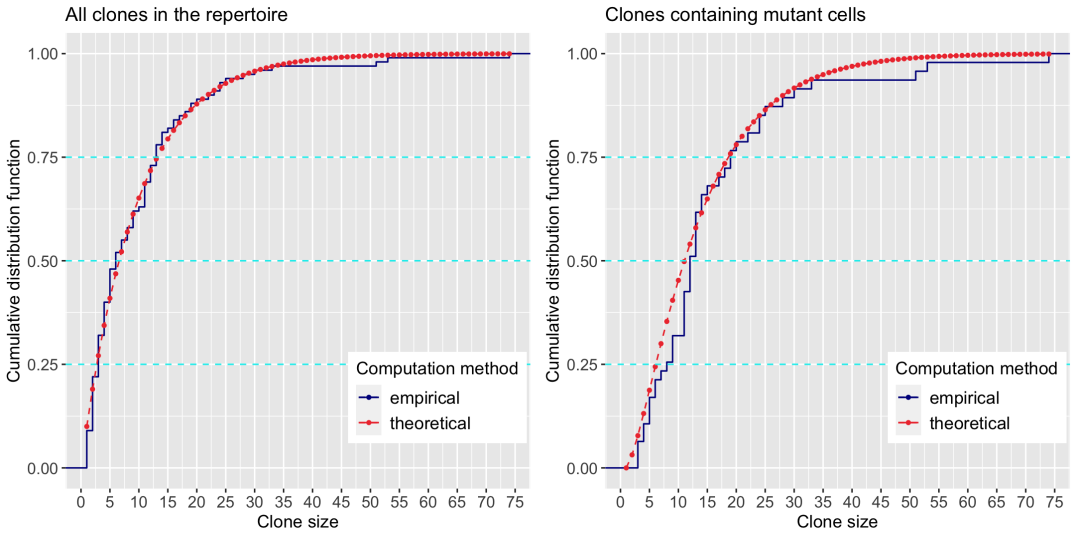

FIG S3. *Comparison of empirical and theoretical cumulative distribution functions (c.d.f.) for simulated repertoire in Fig 5. The theoretical c.d.f is calculated from geometric marginal distribution on clonotype sizes, with parameter* $p = 0.1$ *and mutant frequency* $\theta = 0.05$*. These parameters match the repertoire in Fig 5. The solid dark blue lines are empirical c.d.f's, while dashed red lines are theoretical c.d.f's. Dashed light blue lines are three quantiles, on which clonotypes with mutant cells are* 1.5 *to* 2 *times greater than clonotypes sampled from the whole repertoire.*

TABLE S1

Median clonotype size in the entire repertoire, $P(N_\sigma = n)$, compared to the selected repertoire, $P(N_\sigma = n|M_\sigma = 1)$, for several parametric settings, when $\theta = 10^{-4}$

	Yule-Simon power-law parameter ρ			
	0.05	0.1	0.15	0.2
median: $P(N = n)$	6.13×10^5	622	64	21
median: $P(N = n\|M = 1)$	$> 10^8$	2.80×10^7	1.04×10^5	6276

TABLE S2

Mutant frequency $P(M_S = 1 \mid \aleph_{\text{cel}}, \aleph_{\text{clo}})$ in blood samples computed for hypothetical repertoires with various $\aleph_{\text{clo}}$ and θ settings. Total cells in the repertoire $\aleph_{\text{cel}}$ is fixed at 10^9.

	mutation frequency θ		
$\aleph_{\text{clo}}$	5×10^{-7}	10^{-6}	5×10^{-6}
10^6	7.88×10^{-6}	1.41×10^{-5}	6.93×10^{-5}
5×10^6	7.05×10^{-6}	1.33×10^{-5}	5.56×10^{-5}
10^7	2.71×10^{-6}	7.42×10^{-6}	4.41×10^{-5}

TABLE S3

Repertoire sample characteristics. All samples are from human tissue; PBMC=peripheral blood mononuclear cells; data were downloaded from `https://www.10xgenomics.com/resources/datasets` between June and November, 2022.

Data set	Source	Date published	Cell Ranger
PBMC3	healthy human PBMC	2019-07-24	3.1.0
SC5k	healthy human PBMC	2021-05-25	6.0.1
20k	healthy human PBMC (male, age 30-35)	2021-08-09	6.1.0
10k	healthy human PBMC (female, age 25-30)	2022-03-29	6.1.2
Controller	healthy human PBMC (male, age 30-35)	2021-08-09	6.1.0
Melanoma	dissociated tumor cells from primary human melanoma sample	2021-05-25	6.0.1
Lung	fresh surgical resection of squamous non-small cell lung carcinoma	2021-01-07	5.0.0

Dataset	Genome-aligned gene expression BAM file names
PBMC3	`vdj_nextgem_hs_pbmc3_possorted_genome_bam.bam`
SC5K	`SC5v2_humanPBMCs_5Kcells_Connect_single_channel_SC5v2_humanPBMCs_5Kcells_Connect_single_channel_count_sample_alignments.bam`
20K	`20k_PBMC_5pv2_HT_nextgem_Chromium_X_20k_PBMC_5pv2_HT_nextgem_Chromium_X_count_sample_alignments.bam`
10K	`10k_PBMC_5pv2_nextgem_Chromium_X_intron_10k_PBMC_5pv2_nextgem_Chromium_X_intron_count_sample_alignments.bam`
Controller	`10k_PBMC_5pv2_nextgem_Chromium_Controller_10k_PBMC_5pv2_nextgem_Chromium_Controller_count_sample_alignments.bam`
Melanoma	`SC5v2_Melanoma_5Kcells_Connect_single_channel_SC5v2_Melanoma_5Kcells_Connect_single_channel_count_sample_alignments.bam`
Lung	`vdj_v1_hs_nsclc_multi_5gex_t_b_count_possorted_genome_bam.bam`

Data set	# cells	# T cells	RNA-seq coverage		# SNVs		# clonotypes
			mean	sd	primary	filtered	
PBMC3	11715	6505	24.21	205.72	178456	282	5992
SC5K	4190	1865	13.57	137.83	96942	171	1700
20K	18470	9392	17.98	175.18	121414	157	8728
10K	9780	4642	12.88	151.69	44482	74	4373
Controller	9302	4562	13.46	146.05	69942	85	4282
Melanoma	4680	1434	14.56	159.87	66110	90	1201
Lung	7092	1857	33.13	260.68	319879	750	1482

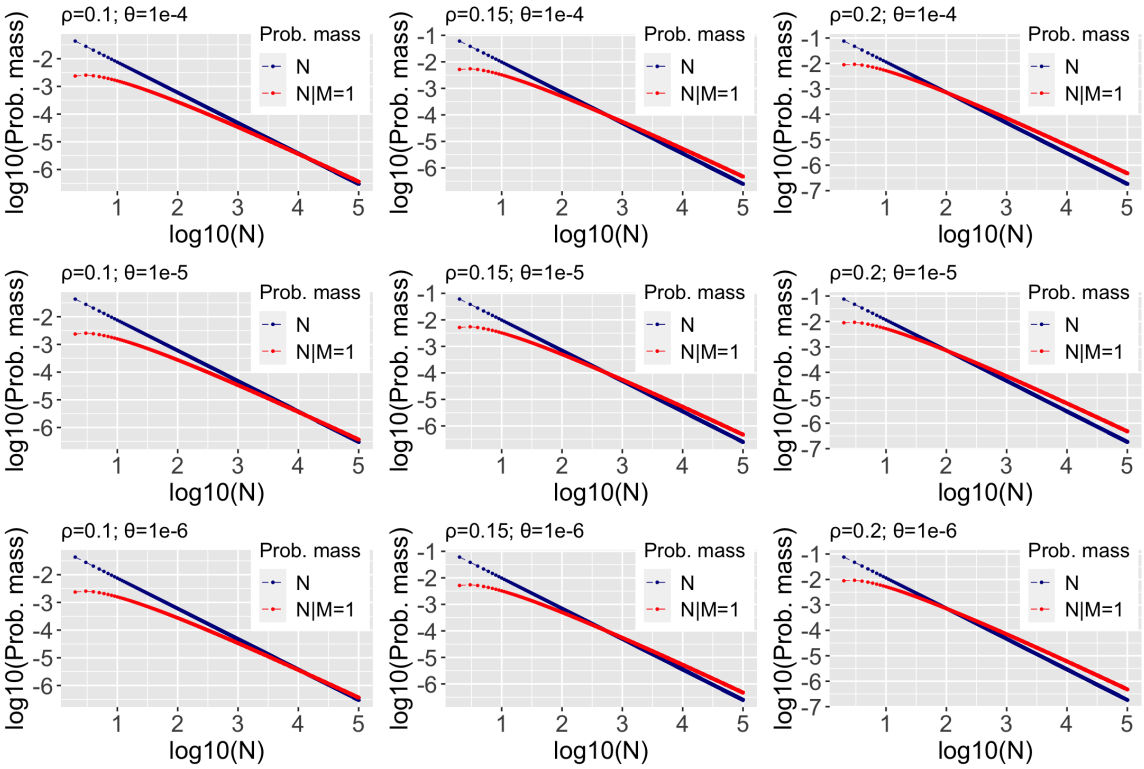

FIG S4. $P(N_\sigma = n | M_\sigma = 1)$ *(red) for various power laws* $P(N_\sigma = n)$ *(blue), on the double log scale.*

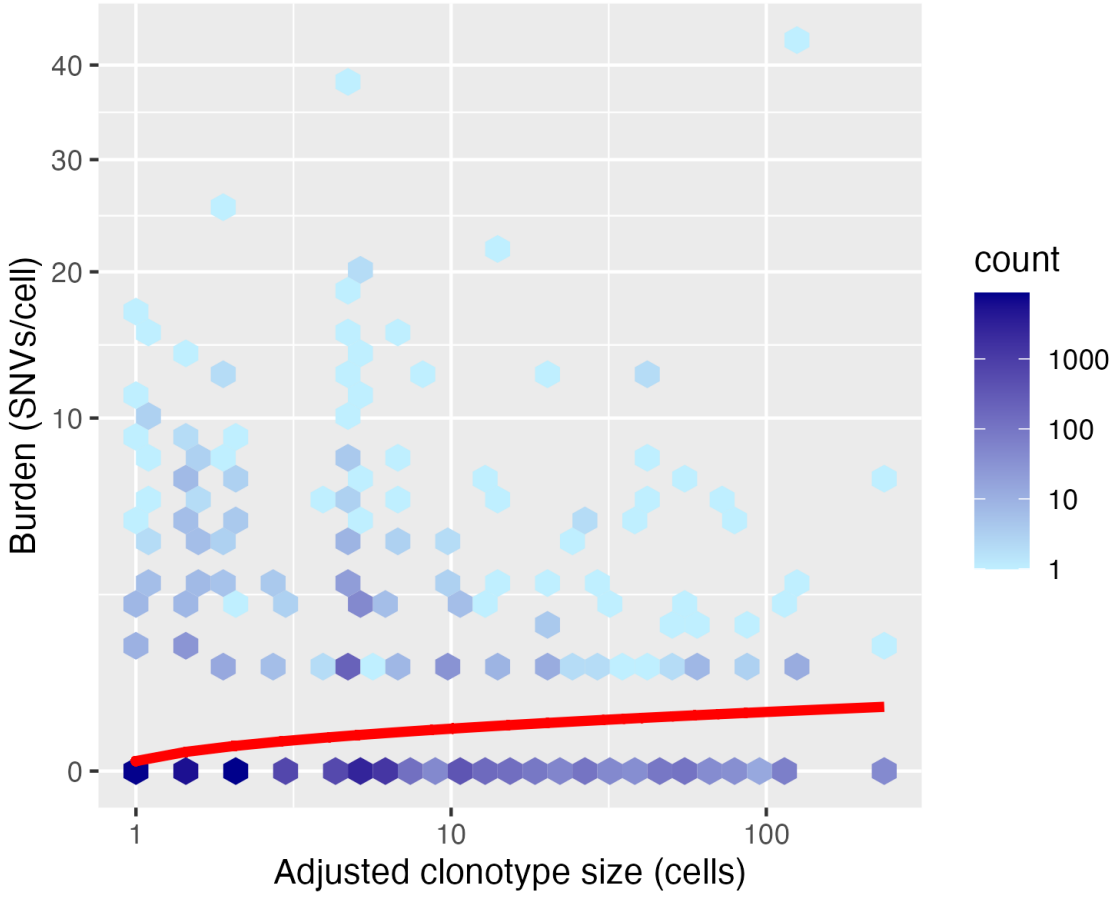

FIG S5. *Each of 30257 T cells from 7 repertoires is associated with a somatic burden (vertical) and also a clonotype size (horizontal), the latter of which is adjusted in an effort to normalize repertoire samples. The red curve shows the estimated effect on expected burden of the logarithm of clonotype size, as determined by a linear model fit ($\hat{\beta} = 0.6$ SNVs per unit increase in* log *clonotype size). Statistical significance of the estimated slope was assessed by a stratified randomization, which shuffled cells between clonotypes but within repertoires (permutation p-value* 0.02 *with* $B = 10^4$ *shuffles). Though statistically significant, the result is not fully resistant; for example the cells in large clonotypes have very high leverage; dropping the cells with adjusted clonotype size greater than 100, for example, leads to an insignificant permutation p-value. The adjusted log size is log clonotype size minus log of repertoire size plus log of largest repertoire size.*