SURROGATE SELECTION OVERSAMPLES EXPANDED T CELL CLONOTYPES

BY PENG YU[1,a], YUMIN LIAN[2], ELLIOT XIE[3], CINDY L. ZULEGER[4,5], RICHARD J. ALBERTINI[6], MARK R. ALBERTINI[4,5,7,c] AND MICHAEL A. NEWTON[1,3,5,b]

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

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

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

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

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

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

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

Surrogate selection is an experimental design that without sequencing any DNA can restrict a sample of cells to those carrying certain genomic mutations. In immunological disease studies, this design may provide a relatively easy approach to enrich a lymphocyte sample with cells relevant to the disease response because the emergence of neutral mutations associates with the proliferation history of clonal subpopulations. A statistical analysis of clonotype sizes provides a structured, quantitative perspective on this useful property of surrogate selection. Our model 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

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

the complexity of a typical T cell population and the potential for an improved understanding 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. Responses among similar patients may vary from morbid toxicity to full recovery (e.g., Ganesan and Mehnert, 2020; Shum, Larkin and Turajlic, 2022), and it may be useful to identify tumor-reactive T cell receptors that could inform therapy (e.g., Pétremand et al., 2024).

Surrogate selection is an experimental design strategy. It restricts a sample of T cells to cells whose somatic ancestors in the study participant had acquired and then transmitted specific, selectable mutations. Selection assays based on mutations of the hypoxanthine-guanine phosphoribosyltransferase (HPRT) gene are the 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 a 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. It remains to be confirmed in clinical settings, but there is potential for surrogate selection to support the monitoring of patients in early phases of immunotherapy by providing a window into their T cell response.

The idea that surrogate selection can enrich for clonally amplified T cells has provided a rationale in many studies, though quantitative treatments of this 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 depends on the biological consequences of *in vivo* clonal proliferation to enrich for activated T cells in individuals with an ongoing immunological response to disease. This enrichment effect is complicated by the enormous complexity of T cell populations and relies on statistical properties of the assemblage of dynamically varying clone sizes. Resolving these complications will enhance our understanding 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 also 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 also 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). This repertoire, which is central to the adaptive immune system, is a union of clonotypes and is the cell population being sampled during data collection. The number of T cells in each clonotype within the repertoire 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). The clonotype-defining TCR affects clonotype size fluctuations, most notably by inducing cell division when the TCR productively interacts with cognate antigen in the presence of appropriate co-stimulatory molecules. Of interest in disease studies are antigens from proteins produced abnormally within a growing cancer, for example, or antigens from normal proteins recognized by a defective immune response in auto-immune disorders. 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. TCR 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. Notably, no explicit DNA sequencing is involved. 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, and thus may be useful in monitoring response to immunotherapy, for example. 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 TCR 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, TCR 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 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 the number of edges along the path from the leaf node to the first cell division; i.e., it is the depth of the leaf node in that reduced tree, assuming $n \geq 2$. For a cell randomly sampled from the clonotype, let D 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 is

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

which is the formulation presented in (Mahmoud, 1992, Page 71-74), Eq. (2.4).[1] Here $\langle x\rangle_{n-1} = x(x+1)(x+2)\cdots(x+n-2)$ is the rising factorial, which is conveniently ex-

[1] In (Mahmoud, 1992, Eq 2.4), a binary tree is assumed to contain n internal nodes and thus $n+1$ 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. We ask that $n \geq 2$ and $1 \leq D \leq n-1$. (In case $n = 1$, no divisions have happened so $D = 0$ w.p. 1.)

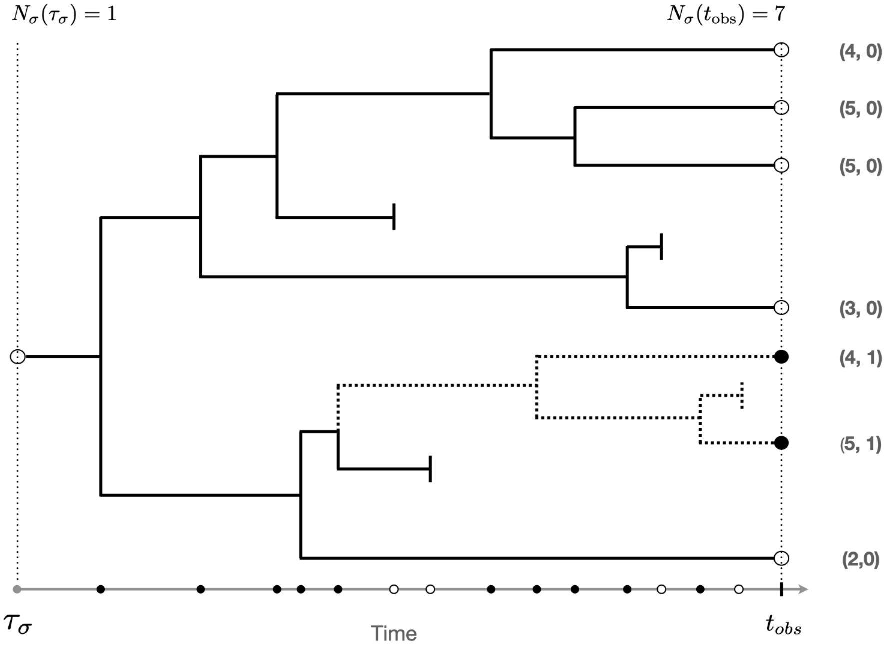

Fig 1: *Tree formed by a developing clonotype σ containing $N_\sigma(t_{\text{obs}}) = 7$ extant cells at time t_{obs} after the initial cell was released from the thymus at time τ_σ. Event times along the timeline indicate the 10 birth events (solid circles) and 4 death events (open circles). The fifth birth event gives rise to a mutant cell, which is ancestral to 2 cells at t_{obs} (dashed). Annotated on the right for each cell is the pair (D, M) recording the number of postthymic divisions leading to that cell and its mutation status, would that cell be sampled at t_{obs}.*

pressed 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 conditional expectation and variance of D, given $N_\sigma(t_{\text{obs}}) = n$, are also available, with both well approximated by twice the natural logarithm of n, and that as n increases, $\{D - 2\log(n)\} / \sqrt{2\log(n)}$ 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 $D|N_\sigma(t_{\text{obs}}) = n$ 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 is,

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

where S gives the unsigned Stirling number of the first kind (e.g., Lynch, 1965; Steel, 2024).

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

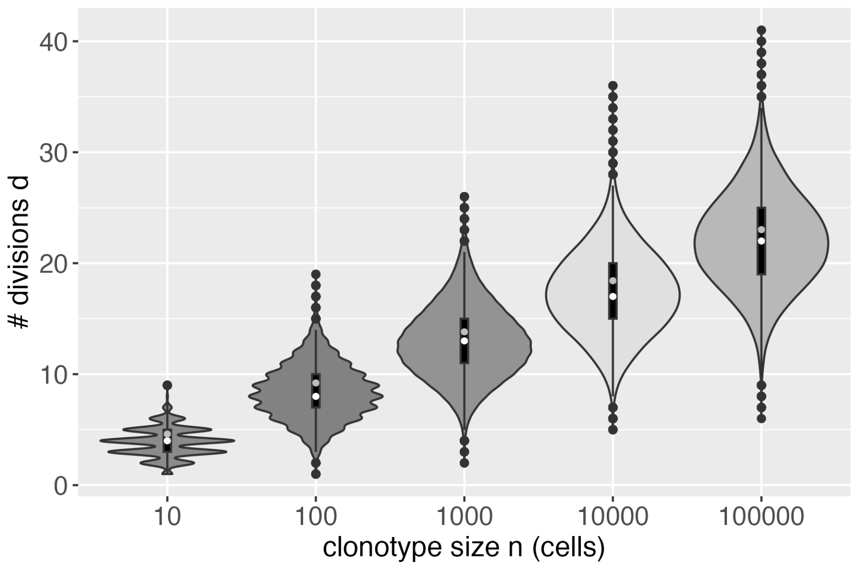

Fig 2: *Proliferation effect: Shown are violin plots of the division number D for cells in randomly developed binary trees, having various sizes, n, at observation time. We used* **R** *packages* ***ape****, to simulate Yule trees, and* ***adephylo****, to count divisions (Paradis and Schliep, 2019;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

a single daughter cell, that are irreversible, and that occur independently across cell divisions. 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. Other mechanisms may induce mutations in both daughter cells or occur separately from mitosis (e.g., Abascal et al., 2021). Though statistical formulations may be adapted to these cases (e.g., Kendall, 1960; Roshan, Jones and Greenman, 2014), we emphasize one specific mechanism for definiteness; robustness of our findings to certain variations in this mechanism are confirmed in Section 5.

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

$$P\{M = 1 \mid D = d, N_\sigma(t_{\text{obs}}) = n\} = 1 - (1-\theta)^d \tag{3}$$

where D 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 and $N_\sigma(t_{\text{obs}})$ are conditionally independent given D. Our first finding concerns the rate of mutant genotype in clonotypes of a given size, and is obtained by marginalizing the distribution of D. Define $\psi_n := P\{M = 1 | N_\sigma(t_{\text{obs}}) = n\}$, and note that for neutral mutations and a Yule tree model, $\psi_1 = 0$, and for $n \geq 2$,

$$\begin{aligned}\psi_n &= \sum_{d=1}^{n-1} P(M = 1 \mid D = d)\, P\{D = d \mid N_\sigma(t_{\text{obs}}) = n\} \\ &= \sum_{d=1}^{n-1} \left\{1 - (1-\theta)^d\right\} P\{D = d \mid N_\sigma(t_{\text{obs}}) = n\}\end{aligned}$$

$$
\begin{aligned}
&= 1 - G_n(1-\theta) \\
&= 1 - \frac{\Gamma(n+1-2\theta)}{\Gamma(n+1)\Gamma(2-2\theta)} \approx 1 - \frac{1}{n^{2\theta}\,\Gamma(2-2\theta)}, \qquad (4)
\end{aligned}
$$

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 times 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's rule.* The development so far has emphasized probabilities that condition on clonotype size. The stochastic evolution of clonotype σ over time induces a distribution on clonotype size at observation time, which we layer in next. For example, the linear pure-birth model leads to the Geometric$\{\exp(-\lambda_\sigma t_{\text{obs}})\}$ distribution,

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

where λ_σ is the birth rate (rate of cell division). Further, compounding over λ_σ gives the Yule-Simon law with parameter ρ, which is inversely proportional to the expectation of λ_σ (Huillet, 2020; Yu et al., 2025).

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

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:

$$
\begin{aligned}
P\{N_\sigma(t_{\text{obs}}) = n \mid M = 1\} &= \frac{P\{M = 1 \mid N_\sigma(t_{\text{obs}}) = n\}\; P\{N_\sigma(t_{\text{obs}}) = n\}}{P(M=1)} \qquad (7) \\
&= \frac{P\{N_\sigma(t_{\text{obs}}) = n\}}{P(M=1)} \left\{1 - \frac{\Gamma(n+1-2\theta)}{\Gamma(n+1)\Gamma(2-2\theta)}\right\}.
\end{aligned}
$$

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 $M = 1$, the median clonotype size shifts up to 8139 cells. This effect is not limited to the marginal Geometric law. Figure 4 shows the result for a Logarithmic distribution (p.m.f. proportional to p^n/n)

and a Yule-Simon law (6), respectively. The enrichment phenomenon holds for any distribution on clonotype size as long as the growth dynamics provide for full support so that all conditional probabilities are well defined. Summarizing the findings for a single developing clonotype, we have:

PROPOSITION 1. *Suppose that 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, that σ began with a single non-mutant cell, and that $P\{N_\sigma(t_{\rm obs})=n\}>0$ for $n\geq 1$. If M indicates that a randomly sampled cell from σ at time $t_{\rm obs}$ is mutant, then the enrichment ratio $\phi_n := P\{N_\sigma(t_{\rm obs})=n \mid M=1\}/P\{N_\sigma(t_{\rm obs})=n\}$ is:*

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

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

Two immediate corollaries assure that: (1) there exists a crossover point $n_{\rm cross}$ with $\phi_n<1$ when $n<n_{\rm cross}$ and $\phi_n>1$ when $n>n_{\rm cross}$, 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: $P\{N_\sigma(t_{\rm obs})\geq n \mid M=1\}\geq P\{N_\sigma(t_{\rm obs})\geq n\}$ (see Pfanzagl, 1964). Among other things, it also follows that the conditional distribution of $N_\sigma(t_{\rm obs})$ given $M=1$ 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.

2.5. *Beyond pure birth.* Relaxing the no-cell-death assumption makes quantifying enrichment more difficult. Explicit calculations show that conditioning on $M=1$ does not necessarily enrich for larger clonotypes. A highly stylized example (Yu et al., 2025) captures features of clonal expansion followed by rapid clonal decline. The intuition is that having sampled a mutant cell, we learn that its containing clonotype is relatively old rather than being relatively large; these two features are equivalent in the pure-birth model. Notwithstanding this counterexample, we find that conditioning on mutation of a sampled cell does enrich for larger clonotypes in a class of well-behaved birth-death processes.

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. It is important for our method of proof that in either case this random selection is uniform among the cells within the clonotype (i.e., neutral steps, e.g., Steel, 2024). By time t, the clonotype size is $N_\sigma(t)=1+\sum_{i=1}^{I(t)}A_i$ where $\tau_{I(t)}\leq t<\tau_{I(t)+1}$ and $I(t)=\max\{j:\tau_j\leq t\}$. We suppose that $N_\sigma(t)$ is not explosive, and thus only a finite number of τ_j's can occur in any finite time interval. We suppose that A_i is conditionally independent of $(\tau_1,\tau_2,\cdots,\tau_i)$ given $A_1,A_2,\cdots,A_{i-1}$, so that the discrete clonal history becomes separable from timing issues. 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). So that conditional probabilities are well defined, we ask that clonotype dynamics assure full support, i.e., that $P\{N_\sigma(t_{\rm obs})=n\}>0$ for all integers $n\geq 1$. As in the pure-birth case, a mutation may

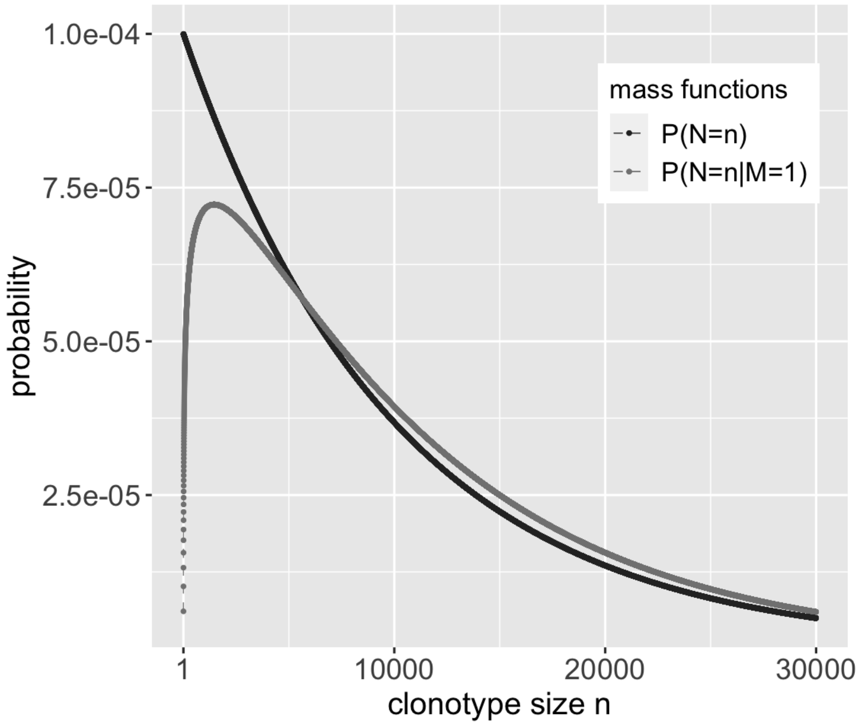

Fig 3: $P\{N_\sigma(t_{\text{obs}}) = n \mid M = 1\}$ *(grey) when the marginal distribution (black) is a Geometric distribution with parameter* $e^{-\lambda t_{\text{obs}}} = 10^{-4}$ *and the mutation frequency* $\theta = 10^{-6}$. *The crossover point* n_{cross} *is* 5624 *cells.*

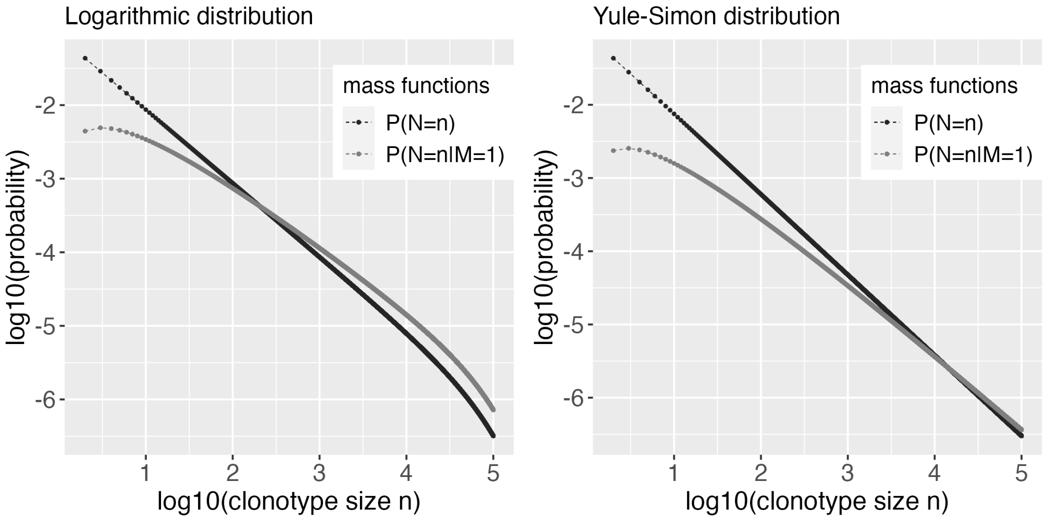

Fig 4: $P\{N_\sigma(t_{\text{obs}}) = n \mid M = 1\}$ *(grey) when the marginal clonotype size distribution (black) is a Logarithmic distribution (left) or a Yule-Simon distribution (right), with parameters* $p = 1 - 10^{-5}$ *for Logarithmic distribution and* $\rho = 0.1$ *for Yule-Simon distribution. 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.*

arise in one daughter of a non-mutant cell in case $A_i = 1$; mutations occur with probability θ per daughter cell, independently of all other properties of the clonotype up to that time (i.e., neutral mutations). Both daughters of a dividing mutant cell are mutant. For a cell sampled randomly from the clonotype at t_{obs}, let M indicate its mutation status, and introduce the conditional mutant frequency $\Psi(a_1, a_2, \cdots, a_i) := P\{M = 1 \mid \mathcal{A}_i,\, I(t_{\text{obs}}) = i\}$, where $\mathcal{A}_i = \cap_{j=1}^{i}(A_j = a_j)$ tracks the specific birth-death sequence. Obviously we cannot sample a cell from an empty clonotype, so we furthermore condition on non-extinction, i.e. $\nu_i \geq 1$ 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 the birth-death process defined above, $Z_i := \Psi(A_1, A_2, \cdots, A_i)$ is non-decreasing in i, converges almost surely to* 1*, and $E(Z_i) = P\{M = 1 \mid I(t_{\text{obs}}) = i\}$ converges to 1. Further, $P\{M = 1 \mid N_\sigma(t_{\text{obs}}) = n\} \geq \psi_n$ for all sizes n, $0 < P(M = 1) < 1$, and the enrichment ratio*

$$\frac{P\{N_\sigma(t_{\text{obs}}) = n | M = 1\}}{P\{N_\sigma(t_{\text{obs}}) = n\}} \longrightarrow \frac{1}{P(M=1)} > 1 \qquad \text{as } n \longrightarrow \infty.$$

Many models meet the requirements of Proposition 2. For example marginal to non-extinction, the linear birth-death process has the A_i's i.i.d., with $P(A_i = 1) = \lambda/(\lambda + \mu)$ for birth rate $\lambda > 0$ and death rate $\mu \geq 0$. It is well known that extinction is almost sure when $\lambda \leq \mu$, but also that extinction occurs with probability μ/λ as long as $\lambda > \mu$ (e.g., Grimmett and Stirzaker, 2001, pg 272). The regularity conditions hardly limit the shape of clonotype-size distributions; they simply assure that $N_\sigma(t_{\text{obs}})$ does not collapse to zero or explode to infinity, and that conditioning events have positive probability. Proposition 2 means that conditioning on mutant status does enrich for larger clonotypes, thus extending Proposition 1 to a broader class of birth-death processes.

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. In adult humans, $\aleph_{\text{cel}}$ and $\aleph_{\text{clo}}$ may be on the order of 10^{11} and 10^{8}, respectively. Considering the 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. 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,

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

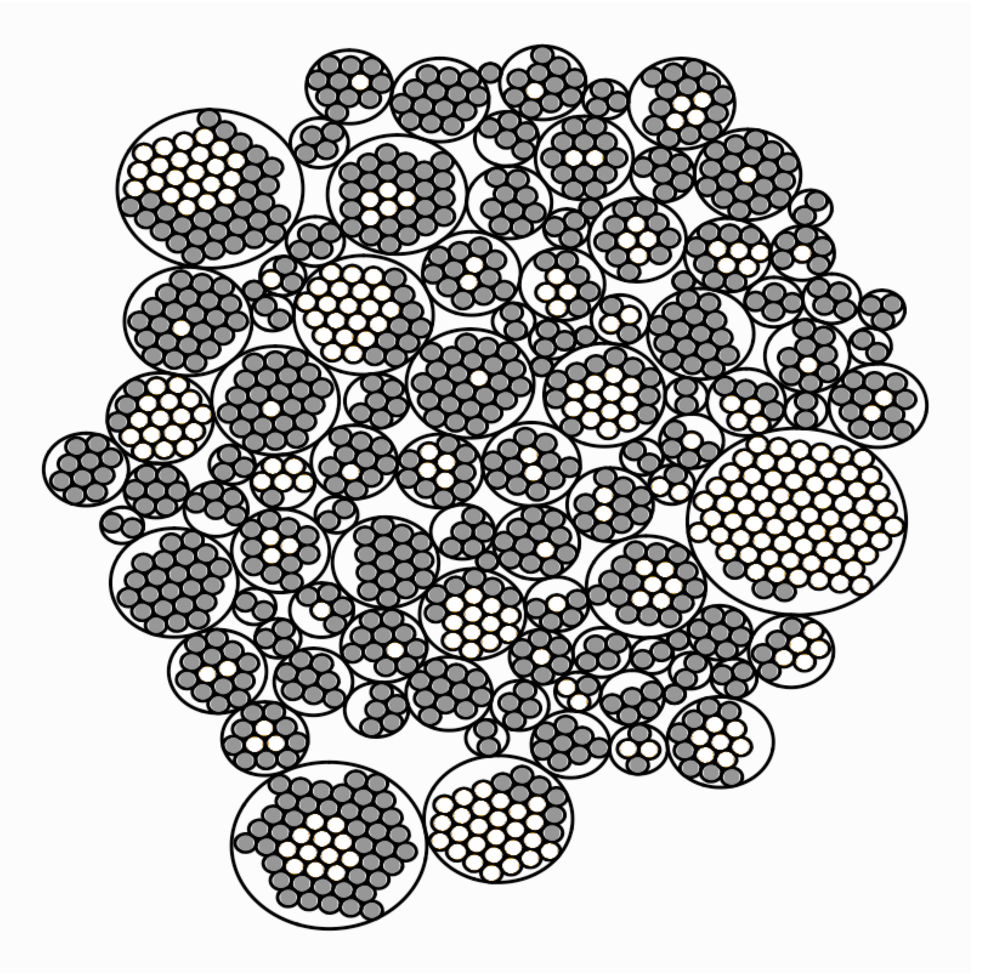

Fig 5: *Simulated repertoire of* $\aleph_{\text{cel}} = 1000$ *cells comprising* $\aleph_{\text{clo}} = 100$ *non-empty clonotypes (encasing circles). The 287 mutant cells are white and the remaining 713 wild-type cells are grey, giving a realized mutant frequency 0.287. As predicted mathematically, 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* $\theta = 0.05$.

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 cell from the repertoire. Conditioning throughout on $\aleph_{\text{cel}}$ and using exchangeability, we have

$$\begin{aligned} P(N_S = n) &= \sum_{\sigma \in \mathcal{S}} P(N_S = n, S = \sigma) = \sum_{\sigma \in \mathcal{S}} P(N_\sigma = n, S = \sigma) \\ &= \sum_{\sigma \in \mathcal{S}} P(S = \sigma \mid N_\sigma = n) \, P(N_\sigma = n) = \sum_{\sigma \in \mathcal{S}} \binom{n}{\aleph_{\text{cel}}} P(N_\sigma = n) \\ &= n P(N_\sigma = n) \binom{\aleph_{\text{clo}}}{\aleph_{\text{cel}}} \qquad \text{for any } \sigma \in \mathcal{S}. \end{aligned} \tag{9}$$

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 than sampling a cell uniformly at random from a randomly developing clonotype. One consequence of (9) is that N_S is stochastically larger than N_σ (Arratia, Goldstein and Kochman, 2019; Yu et al., 2025). 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):

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

where the system-defining parameters $\alpha > 0$ and $p \in (0,1)$ reflect properties 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, conditioning on $\aleph_{\text{cel}} = \sum_j N_{\sigma_j}$ to eliminate p, and setting $\alpha = 1$ gives 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 p^{n_j}/n_j; 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 S5, Yu et al., 2025). 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, Yu et al., 2025); 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 S1. Here, we exercise these reference distributions primarily to explore single versus multi-clonal models, but we recognize they would be useful building blocks in model-based analysis of clonotype-size data. Related to Section 2.5, these reference distributions also emerge from considerations of dynamics of birth-death processes (Dessalles, D'orsogna and Chou, 2018).

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:

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

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

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

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), if each individual clonotype evolves to its size at t_{obs} according to the dynamics in Proposition 1 or Proposition 2, and if M and S are the mutation status and clonotype identifier of a cell drawn randomly from the full repertoire, then the enrichment ratio $P(N_S=n \mid M=1)/P(N_S=n)$ 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. From the perspective of experimental design, Proposition 3 affirms and quantifies the sampling effects of surrogate selection.

3.4. *Mutant Frequency.* At least in the absence of cell death, a random cell from the repertoire is more likely to be mutant than a random cell from any specific, randomly developing clonotype: $P(M=1) \geq P(M=1 \mid S=\sigma)$, which we confirm in the Appendix D by a calculation similar to (9). This mutant frequency $P(M=1)$ 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. Some numerical comparisons are provided in Table S2 and Fig S2 (Yu et al., 2025).

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:

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

Using probability generating 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,

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

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

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

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

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

where ${}_2F_1(a, b; c; z)$ is the Gaussian hypergeometric function. 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 $\tilde{\epsilon} = n_{\text{samp}} / \{\aleph_{\text{cel}} P(M=1)\}$, 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

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

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 $\mathrm{Logarithmic}(p)$*, then the expected diversity is:*

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

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 the expected diversity of a repertoire sample. The result extends to more general distributions by mixing. For example, if conditional upon γ_0 the clonotype sizes are $\mathrm{Geometric}(\gamma_0)$, and if $\gamma_0 = \exp(-W)$ for $W \sim \mathrm{Exp}(\rho)$, then marginally the clonotype size is Yule-Simon distributed with parameter ρ, and the expected diversity bound carries through the expectation: $E\left\{E\left(\mathcal{D} - \tilde{\mathcal{D}} \mid \gamma_0\right)\right\} > 0$.

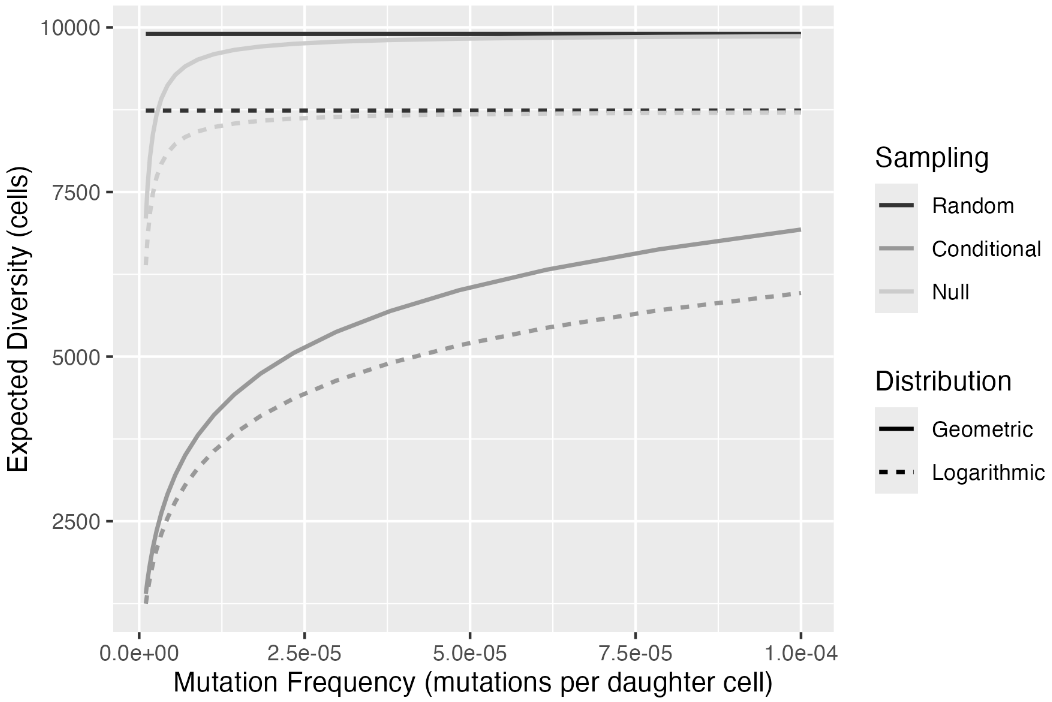

Fig 6: *Expected sample diversity (vertical) at various mutation frequencies* θ *(horizontal) and for different marginal distributions of clonotype size (Geometric or Logarithmic) and considering sampling the whole repertoire (random), sampling the mutant subset (conditional), and a null version sampling the mutant fraction where mutant status is independent of clonotype size (null). Calculations use a repertoire of size* $\aleph_{\mathrm{cel}} = 10^9$ *cells and* $\aleph_{\mathrm{clo}} = 10^6$ *clones, and a sampling fraction* $\epsilon = 10^{-5}$*. The Geometric parameter is set* $\gamma_0 = \aleph_{\mathrm{clo}}/\aleph_{\mathrm{cel}}$ *and the Logarithmic parameter* $p = 0.9998779$ *solves* $-(1-p)\log(1-p) = p\gamma_0$*, following (Watterson, 1974, eq. 3.1). Expected sample diversity is always lower in the mutant fraction than in random draws from the whole repertoire, in accordance with Proposition 4.*

A skeptic uncertain about the extent of surrogate selection's enrichment effect would point out that sample diversity could be lower in the mutant fraction owing only to the effect of reducing the population size. In sampling a repertoire fraction $\tilde{\epsilon}$ to recover n_{samp} mutant cells, we might have very few cells to draw from when $P(M=1)$ is very small, and this setting would result in lower sample diversity whether or not the event $M=1$ enriches for larger clonotypes. Having mutation independent of N_σ provides a ready *null* version of Proposition 4, details of which are in Yu et al. (2025), Appendix E. Figure 6 provides a numerical illustration, and shows that expected sample diversity is lower when sampling the mutant fraction, above and beyond the drop expected by reduced-population-size effects alone.

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_g relative to its prethymic state. We are thinking

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

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_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, $P\left(M_g^m = 1 \mid N_S = n\right) = \psi_n(\theta_g^m)$ for locus-specific mutation frequency θ_g^m, 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 $L^m = \sum_{g\in\mathcal{G}^m} M_g^m$.

PROPOSITION 5. *If clonotypes satisfy the regularity conditions in Proposition 1, if clonotype sizes are exchangeable as in (8), and if* $\lambda^m = \sum_{g\in\mathcal{G}^m}\theta_g^m \longrightarrow \lambda$ *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 postthymic somatic mutations in a T cell (i.e., the expected somatic burden) 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 it 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. Yu et al. (2025), 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 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, first we repurpose the RNA-seq reads to report on underlying somatic mutations that 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 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 postthymic 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 sce-SNVs.

Figure S3 (Yu et al., 2025) 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 (Figure 7). 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.

Cell-function data provide an intriguing validation of the association between somatic burden and clonotype size. Recall that within the T cell repertoire, cells specialize towards a variety of distinct functional subtypes that are distinguished by features of the single-cell transcriptional profile (e.g., Andreatta et al., 2021; Ianevski, Giri and Aittokallio, 2022). These include naive cells which have yet to have been activated by antigen, and various effector, memory, and exhausted subtypes. Figure S4 shows a 2-dimensional embedding of 22424 T cells according to their single-cell transcriptional profiles; these constitute the subset of cells analyzed above for which a confident functional T cell subtype classification was available. Cells thus have a subtype label, in addition to a somatic burden score and clonotype size score; Figure 8 summarizes two of the interesting pairwise marginal distributions, and provides an empirical assessment that is fully in line with our theoretical development and our understanding of T cell biology. Sampled T cells are predominantly naive, with predominantly singleton TCRs and without any somatic variants. Further, as confirmed by formal hypothesis tests (Appendix H), naive T cells correspond to smaller clonotypes and have less somatic burden than non-naive T cells.

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.

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

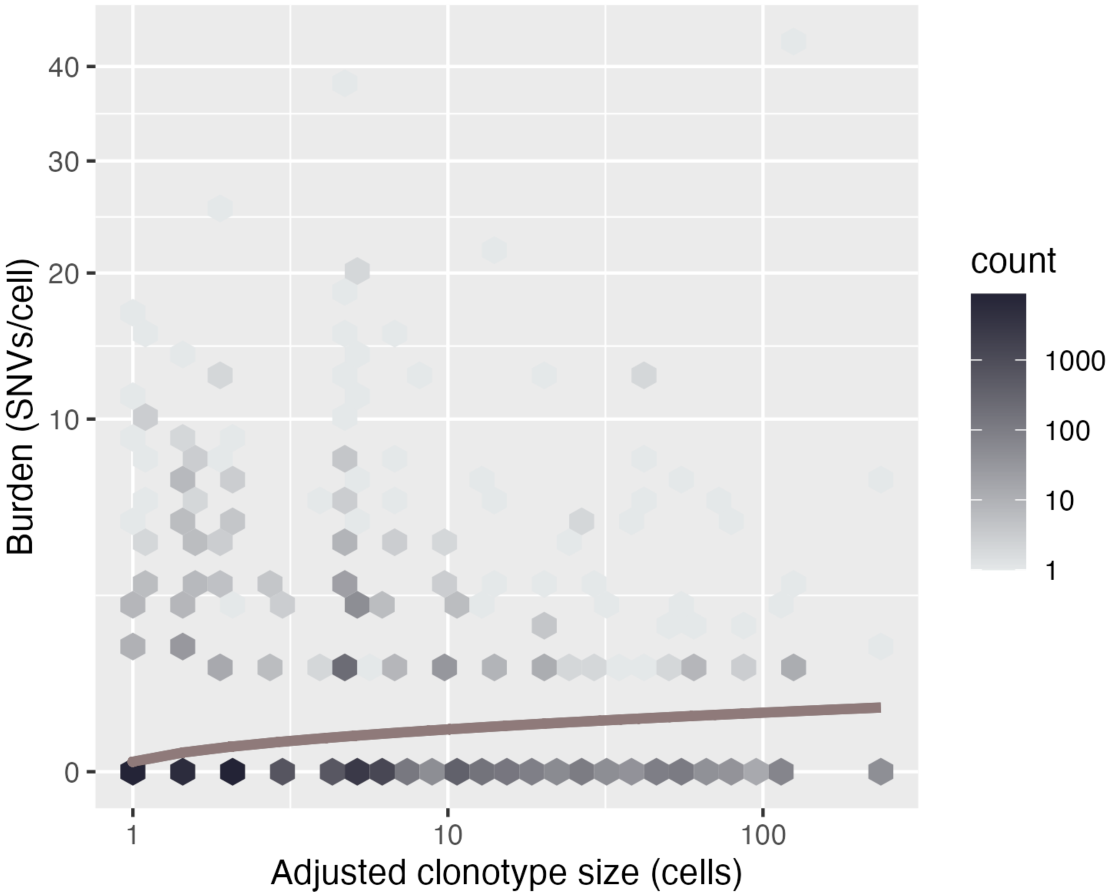

Fig 7: *Each of 30257 T cells from 7 repertoires is associated with a somatic burden (vertical) and also a clonotype size (horizontal), the latter of which is adjusted in an effort to normalize repertoire samples. The curve shows the estimated effect on expected burden of the logarithm of clonotype size, 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.*

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 TCR 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 could readily enrich a lymphocyte sample for T cells responsive to disease-relevant antigens would have a variety of applications; for instance, they could be useful in monitoring a patient's response to immunotherapy. The present work provides a statistical basis to the use of surrogate selec-

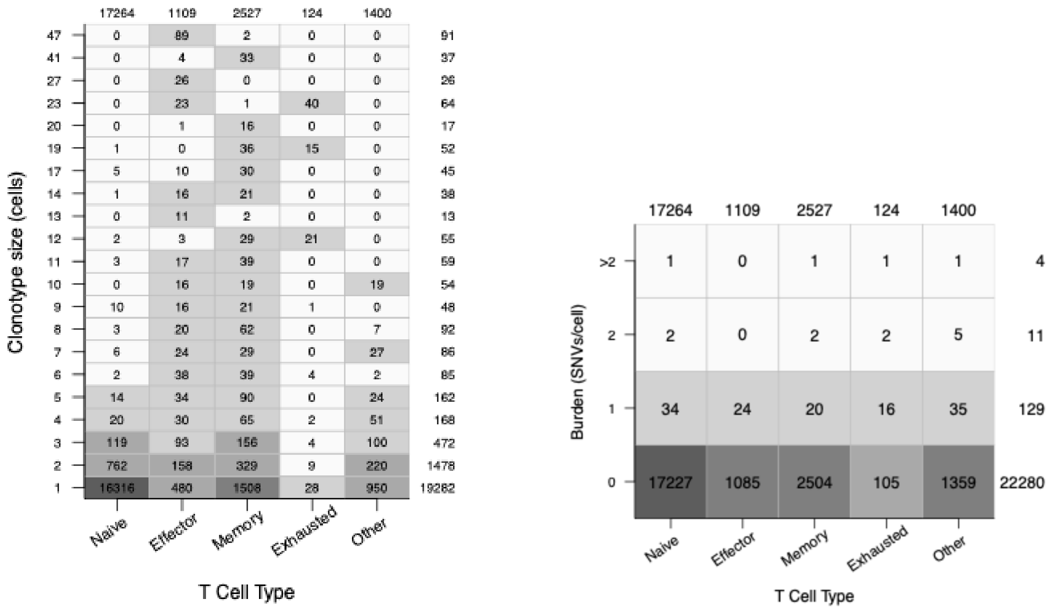

Fig 8: *Over the 22424 cells from 7 T cell repertoires for which scRNA-seq data provide a confident cell type call, shown are empirical joint distributions of cell type (horizontal) and clonotype size (vertical, left panel) or somatic burden (vertical, right panel). Grey scale is by bins at cut points* 10, 10^2, 10^3*, and* 10^4*, and marginal counts are shown on outer edges.*

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 patients; Subjects 26, 29, 30, 32 are healthy donors. Subjects are sorted by the number of sequenced TCRs.*

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

tion, 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. Extensive efforts by others have advanced birth-death processes for immunodynamics (e.g., Roshan, Jones and Greenman, 2014; Dessalles, D'orsogna and Chou, 2018; Molina-París and Lythe, 2021); to our knowledge, prior work does not examine the sampling effects caused by conditioning on a mutant fraction of the population. However, such examination sheds light on a potentially useful experimental design strategy.

Our theoretical predictions are accompanied by empirical results both from surrogate selection studies and recent single-cell sequencing projects. One take-home message is that we have resolved the sampling phenomenon exemplified in the simulated data of Figure 5. 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. We also report on a repurposing of scRNA-seq data to assess somatic genomic changes and we offer a new somatic burden statistic for T cells. In spite of high noise levels we detect a significant association between clonotype size and somatic burden. Owing to our restricted permutation method, the result is protected from the effects of confounding variables that vary among repertoires, and the result is validated in part by cell lineage calculations (Figure 8). In future efforts, we expect that (1) more reliable genomic DNA sequence data may become available for this purpose, and (2) that better variant calling pipelines could be customized for repertoire sampling (e.g., to recognize the common germline of sampled cells). Improvements in somatic variant calling could foster other statistical calculations to infer properties of clonotype dynamics (e.g., Figure S7, Yu et al. (2025)). They may also improve the detection and removal of prethymic mutations, which are less relevant to immunological surveillance.

Our focus is an applied problem that has not received formal statistical analysis in prior literature. To study the enrichment phenomenon, we place great emphasis on the stochastic aspects of a person's time-bound developing T cell repertoire. Certain modeling features, like exchangeability of clonotype sizes, have little basis in biology, yet they have tremendous practical appeal: for one thing, contemporary immuno-profiling studies report frequency spectra and draw inferences from the shape of these spectra. These frequency spectra are sufficient statistics in exchangeable models, as we remind readers in Section 3.1. While contemporary data sets are relatively large compared historical data, nevertheless the amount of data on individual clonotypes remains very limited, and analysts are not well positioned to entertain non-exchangeable models in many practical settings. Other assumptions, such as neutrality in the identity of cells that die, mutate, or divide, are popular in cell biology; violations are no doubt possible, but the assumptions taken constitute a baseline from which more elaborate models may be derived. Assumptions on the birth-death processes are not parametric; they provide some general constraints on the developing repertoire.

Non-mitotic mutations have not been considered in our development, though they can occur by a variety of mechanisms (e.g., Abascal et al., 2021). Statistically, non-mitotic mutations add background noise to a sampled cell's somatic burden count, L, and elevate a given locus' mutant frequency $P(M = 1)$ above what is induced through mitotic mutation alone. While the history of two different cells could reflect different exposure risks for non-mitotic mutation, we note that all cells in the repertoire at time t_{obs} have experienced the same overall time for non-mitotic mutations to emerge; at least they descend from the conception event, and surely they all descend from some progenitor of the hematopoietic stem cell pool. So, any two cells at t_{obs}, no matter how many differences in their mitotic pasts, will have the same time for accumulation of non-mitotic mutations. If the probability is κ that a non-mitotic mutation arises at a specific locus by time t_{obs}, then actual mutant frequency $P(M=1) = 1-(1-\kappa)P(M_{\mathrm{mitotic}}=0)$. From this we find the enrichment ratio ϕ_n retains the monotonicity properties as confirmed in the purely mitotic cases in Propositions 1, 2, and 3, but the ultimate enrichment is reduced by an amount related to the noise level κ.

The joint distribution of clonotype sizes, discussed briefly in Section (3.2) warrants further study in the context of T cell repertoires. We note, for example, that the exchangeable formulation (8) associated with Logarithmic marginals induces the Ewens sampling formula as the distribution on sufficient statistics (the counts of counts), which is of fundamental importance in other domains (e.g., Crane, 2016; Tavaré, 2021). The clonal enrichment phenomenon is not restricted to particular joint distributions, however coupling our calculations to specific repertoire models (e.g., Böttcher, Wald and Chou, 2023) may enable more quantitative assessment of the surrogate selection effect. Studies of immunosenescence, or the age-related deterioration of the immune system, might also benefit from surrogate selection to probe the more actively aging immune components (e.g., Liu et al., 2023). We note too that surrogate selection is a limited tool, in part because a multitude of causes beyond the presence of disease antigens may be stimulating clonotype proliferation. In any case, we hope that our work supports informed statistical analysis of T cell data sets, planning of immunological experiments, and applications to monitoring immune response.

SUPPLEMENTARY MATERIAL

Supplementary Calculations (DOI: ** SurrogateSelectionSupplement.pdf)
We provide derivations, proofs, and additional modeling elements in support of findings presented above, as well as further details on data preparation and analysis for Section 4. Material in Yu et al. (2025) is organized into nine appendices that refer to appropriate sections of the present paper.

REFERENCES

Abascal, F., Harvey, L. M., Mitchell, E., Lawson, A. R., Lensing, S. V., Ellis, P., Russell, A. J., Alcantara, R. E., Baez-Ortega, A., Wang, Y. et al. (2021). Somatic mutation landscapes at single-molecule resolution. *Nature* **593** 405–410.

Albertini, R. J. (2001). HPRT mutations in humans: biomarkers for mechanistic studies. *Mutation Research/Reviews in Mutation Research* **489** 1-16. https://doi.org/10.1016/S1383-5742(01)00064-3

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.

Andreatta, M., Corria-Osorio, J., Müller, S., Cubas, R., Coukos, G. and Carmona, S. J. (2021). Interpretation of T cell states from single-cell transcriptomics data using reference atlases. *Nature communications* **12** 1–19.

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

Arratia, R., Goldstein, L. and Kochman, F. (2019). Size bias for one and all. *Probability Surveys* **16** 1–61.

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.

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.

Böttcher, L., Wald, S. and Chou, T. (2023). Mathematical Characterization of Private and Public Immune Receptor Sequences. *Bulletin of Mathematical Biology* **85** 102.

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.

CRANE, H. (2016). The ubiquitous Ewens sampling formula. *Statistical science* **31** 1–19.

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.

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.

DESSALLES, R., D'ORSOGNA, M. and CHOU, T. (2018). Exact steady-state distributions of multispecies birth–death–immigration processes: Effects of mutations and carrying capacity on diversity. *Journal of statistical physics* **173** 182–221.

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. https://doi.org/10.1093/bioinformatics/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. https://doi.org/10.1038/s41591-019-0734-6

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. https://doi.org/10.1146/annurev-cancerbio-030419-033604

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

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.

HUILLET, T. E. (2020). On new mechanisms leading to heavy-tailed distributions related to the ones of Yule-Simon. *Indian Journal of Pure and Applied Mathematics* **51** 321–344.

IANEVSKI, A., GIRI, A. K. and AITTOKALLIO, T. (2022). Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. *Nature communications* **13** 1246.

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.

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. https://doi.org/10.1093/bib/bbab160

LIU, Z., LIANG, Q., REN, Y., GUO, C., GE, X., WANG, L., CHENG, Q., LUO, P., ZHANG, Y. and HAN, X. (2023). Immunosenescence: molecular mechanisms and diseases. *Signal transduction and targeted therapy* **8** 200.

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. https://doi.org/10.1038/s41591-021-01623-z

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. https://doi.org/10.1152/advan.00066.2013

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. https://doi.org/10.1016/j.mrrev.2009.12.004

PÉTREMAND, R., CHIFFELLE, J., BOBISSE, S., PEREZ, M. A., SCHMIDT, J., ARNAUD, M., BARRAS, D., LOZANO-RABELLA, M., GENOLET, R., SAUVAGE, C. et al. (2024). Identification of clinically relevant T cell receptors for personalized T cell therapy using combinatorial algorithms. *Nature Biotechnology* 1–6.

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. (2024). Neutral phylogenetic models and their role in tree-based biodiversity measures. *arXiv preprint arXiv:2405.17833*.

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. https://doi.org/10.1112/blms.12537

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. https://doi.org/10.1038/s43018-019-0022-x

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. https://doi.org/10.1038/s41467-021-24343-x

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.

Watterson, G. A. (1974). Models for the logarithmic species abundance distributions. *Theoretical population biology* **6** 217–250.

Yu, P., Lian, Y., Xie, E., Zuleger, C. L., Albertini, R. J., Albertini, M. R. and Newton, M. A. (2025). Supplement to "Surrogate selection oversamples expanded T cell clonotypes".

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. https://doi.org/10.3389/fimmu.2017.01809

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. https://doi.org/10.1016/j.jim.2010.12.007

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.