A control-calibrated E-value for fuzzy TCR sequence search over biologically redundant reference sets

Mikhail Shugay[1,2,3*]

[1]Institute of Translational Medicine, Russian National Medical State University, Moscow, Russia
[2]Department of Genomics of Adaptive Immunity, Shemyakin–Ovchinnikov Institute of Bioorganic Chemistry RAS, Moscow, Russia
[3]Institute of Molecular Biology NAS RA, Erevan, Armenia

seqtree technical appendix

Abstract

We derive a BLAST-style E-value [7, 1] for "hits" returned by fuzzy search over T-cell receptor (TCR) CDR3 sequences, adapted to the defining difficulty of immune repertoires: the reference set is highly *redundant*, and the redundancy is biological (convergent V(D)J recombination, public clones, clonal expansion) rather than statistical noise. The classical Karlin–Altschul theory assumes a database of independent, identically distributed letters; under that null, redundancy-driven near-matches are absurdly significant. We replace the i.i.d.-letter null with an *empirical background* P_0 estimated from a matched control repertoire, retain the Poisson/Gumbel limit superstructure with an explicit non-asymptotic error bound (Chen–Stein / Le Cam [4, 9]), and handle clonal over-dispersion by collapsing to unique clonotypes. The resulting E-value is automatically deflated for hits that the generation process alone explains and is large only for antigen-driven convergence. This puts the TCRNET approach—counting sequence neighbours against a real-world control repertoire, first introduced by Ritvo et al. [12] and formalized as an annotation framework by [11]—on a rigorous, finite-sample footing, and we show the classical Karlin–Altschul E-value is its product-measure, ungapped special case.

1 Introduction: the redundancy problem

Given a query CDR3 q and a target set D (e.g. VDJdb), fuzzy search returns the neighbours of q within a fixed scope/budget θ. We want a significance value for such hits. BLAST answers this for protein search with the Karlin–Altschul E-value

$$E = K\, m\, n\, e^{-\lambda^* S}, \tag{1}$$

where m is the query length and n the total database length (both in residues), S is the alignment score of the hit under a substitution matrix with entries s_{ij}, λ^* is the unique positive root of $\sum_{ij} p_i p_j e^{\lambda^* s_{ij}} = 1$ (the natural scale that turns scores into log-probabilities for i.i.d. letters with background frequencies p_i), and $K > 0$ is a prefactor—the "clumping" or edge-effect constant—fixed by the score distribution. E is the expected number of distinct alignments scoring at least S by chance; the number of such alignments is asymptotically Poisson, so Pr(at least one) $= 1 - e^{-E}$ [7, 8]. The whole construction rests on the database being a string of i.i.d. letters.

*Correspondence: mikhail.shugay@gmail.com

Immune repertoires violate the i.i.d. assumption catastrophically. CDR3s are produced by V(D)J recombination, whose generation probability P_{gen}, first derived and inferred from sequence repertoires by Murugan et al. [10], is sharply non-uniform; convergent recombination makes some sequences enormously over-represented; clonal expansion and public clones create exact and near duplicates. A query in a common, high-P_{gen} region of sequence space has many neighbours *for purely generative reasons*. An i.i.d. null would flag these as wildly significant, which is biologically meaningless. The signal we actually want is the opposite: *more* neighbours than the background generative process predicts, the hallmark of antigen-driven selection.

Our approach: define the null by an empirical background distribution P_0 that carries the generative and baseline-sharing redundancy but no antigen-driven enrichment, estimate the null neighbourhood mass from a matched control repertoire (the user-supplied `isalgo/airr_control` set), and calibrate the E-value against it. This is a rigorous, finite-sample generalization of Karlin–Altschul to a non-i.i.d., biologically structured null, and the statistical formalization of TCRNET-style neighbour counting against a real control [12, 11].

2 Setup and notation

Let Σ be the amino-acid alphabet and $\mathcal{X} = \bigcup_{L\geq 0} \Sigma^L$ the space of CDR3 sequences. The search engine defines, for a query q and budget $\theta \geq 0$, a non-negative score $s_\theta(q, x)$ and a *ball*

$$B_\theta(q) = \{x \in \mathcal{X} : s(q, x) \leq \theta\}, \qquad s(q, q) = 0,\ s \geq 0. \tag{2}$$

The score need not be a metric: with a substitution matrix it is the squared-distance penalty $\text{pen}(a, b) = s_{aa} + s_{bb} - 2s_{ab}$ summed along the optimal alignment (plus gap costs); in unit-cost mode it is an edit count. Both define a legitimate ball. We work with two background laws on $\mathcal{X}$: the *realized-repertoire background* P_0 (what a healthy, unselected repertoire instantiates) and the *generation law* P_{gen} (the V(D)J model). Let

$$\pi_0(q, \theta) = P_0(B_\theta(q)) = \sum_{x \in B_\theta(q)} P_0(x), \qquad \pi_{\text{gen}}(q, \theta) = P_{\text{gen}}(B_\theta(q))\,. \tag{3}$$

A control sample $C = (C_1, \ldots, C_M)$ and a target set D are given. Crucially, all counts are over *distinct clonotypes*: the engine deduplicates hits by reference id, so

$$n_S(q, \theta) = \#\{\, x \in S \text{ distinct} : x \in B_\theta(q)\,\}, \qquad S \in \{C, D\}. \tag{4}$$

Write $N = |D|$ (unique clonotypes; see §5).

Lemma 1 (Scope monotonicity)**.** *The balls nest, $B_\theta(q) \subseteq B_{\theta'}(q)$ for $\theta \leq \theta'$ (since $s \geq 0$ and the cut is by a single threshold). Hence $\pi_0(q, \cdot)$, $n_S(q, \cdot)$, the intensity $\lambda(q, \cdot)$ and the E-value $E(q, \cdot)$ are all non-decreasing in the scope/budget θ, and the closest-hit score $S_{\min}(q)$ of §9 is the smallest θ with $n_D(q, \theta) > 0$. This justifies sweeping θ to trace an E-value curve per query.*

Assumption 1 (Exchangeability under H_0)**.** Under the null, the unique clonotypes of D are exchangeable with marginal law P_0.

Assumption 2 (Independent control draws)**.** The unique clonotypes of C are i.i.d. (or exchangeable) $\sim P_0$.

Assumption 3 (Background match)**.** C and D share the background P_0 (same generation + sampling process, matched chain, species and length composition). All validity is conditional on Assumption 3.

3 Null hypothesis and estimator hierarchy

Definition 1 (Per-query null)**.** $H_0(q)$: the neighbours of q in D arise from P_0 with no antigen-driven excess, i.e. each $x \in D$ satisfies $\mathbb{E}[\mathbf{1}(x \in B_\theta(q))] = \pi_0(q,\theta)$. The alternative $H_1(q)$ posits excess mass $\pi_D(q,\theta) > \pi_0(q,\theta)$.

Lemma 2 (Self-match exclusion / punctured null)**.** *When the query is itself a database member ($q \in D$, as in a VDJdb-vs-VDJdb scan), the count $n_D(q,\theta)$ contains the exact self-match (and any exact duplicates of q), which are* deterministic *identity hits, not random draws from P_0. Including them biases both the observed count and the null. The correct neighbour statistic is the* punctured *count over the distance-positive ball,*

$$n_D^{>}(q,\theta) = \#\{x \in D : 0 < s(q,x) \le \theta\}, \tag{5}$$

with null intensity $\lambda^{>}(q,\theta) = (N - m_q)\,\pi_0^{>}(q,\theta)$, where m_q is the multiplicity of q in D and $\pi_0^{>} = P_0(B_\theta(q)) - P_0(\{x : s(q,x) = 0\})$ removes the point mass at exact matches. The control estimator is punctured identically, $\hat{\pi}^{>} = n_C^{>}(q,\theta)/M$, so the deterministic identity term cancels in the calibrated E-value. (For $q \notin D$ the puncture is vacuous and $n_D^{>} = n_D$.)

Remark 1 (Consistency of the puncture, and when *not* to use it). The puncture is valid *only if applied to both sides*: the E-value $E = (N/M)\,n_C$ estimates $N\pi_0$ for one and the same ball, so dropping the $s = 0$ point mass from the target count requires estimating the punctured mass $\pi_0^{>}$ from the *punctured* control count $n_C^{>}$. Doing so does change the numeric E-value (it shrinks by the removed exact-match mass, $E^{>} = (N/M)n_C^{>} \le E$), but it leaves the *inference* unbiased: the exact-match term is deterministic and enters observed count and null intensity identically, so it carries no signal and its removal neither creates nor destroys significance for the genuine neighbours. Puncturing only one side (target but not control, or vice versa) *does* bias the test and must be avoided.

This exclusion is a **benchmark device**, not a default for applications. In the VDJdb-vs-VDJdb benchmark the queries are drawn from the target, so every query carries a guaranteed trivial self-hit that would otherwise inflate every count uniformly; puncturing removes it. In a real annotation task the query is a *novel* sequence scored against a reference database ($q \notin D$), where an exact database match is the strongest and most informative hit and must be kept. Hence `seqtree.evalues` leaves `exclude_exact=False` by default and the benchmark sets it `True`.

The estimand is the per-query Poisson intensity $\lambda(q,\theta) = N\,\pi_0(q,\theta)$ (read as $\lambda^{>}$ with the puncture of Lemma 2 whenever $q \in D$). Two estimators of π_0 target *different* nulls and must not be conflated.

- **Control / Monte-Carlo (primary):** $\hat{\pi}(q,\theta) = n_C(q,\theta)/M$, unbiased for $P_0(B_\theta(q))$, with $M\hat{\pi} \sim \text{Binomial}(M, \pi_0)$ under Assumption 2. It captures the *realized* background, including public-clone sharing and finite-repertoire convergence.
- **Generation / analytic (cross-check):** $\hat{\pi}_{\text{gen}}(q,\theta) = \sum_{x \in B_\theta(q)} P_{\text{gen}}(x)$, computed by enumerating the (small, for small θ) ball with the engine and weighting by the V(D)J generation probability of the Murugan et al. model [10]. It targets the pure generation null $P_{\text{gen}}(B_\theta(q))$, which omits selection and sampling.

Remark 2 (Selection factor and the thymic correction). P_{gen} is a *pre-selection* law; only a fraction of generated receptors survive thymic and peripheral selection. Elhanati et al. [6] model this with

a per-sequence *selection factor* $Q(\sigma) \geq 0$ on the recombination outcome $\sigma = (\vec{a}, V, J)$, inferred by maximum likelihood, giving the post-selection law

$$P_0(\sigma) = \tfrac{1}{Z}\, Q(\sigma)\, P_{\text{gen}}(\sigma), \qquad Z = \sum_{\sigma} Q(\sigma) P_{\text{gen}}(\sigma) = 1 \;\; (\langle Q \rangle_{P_{\text{gen}}} = 1). \tag{6}$$

The normalization $\langle Q \rangle = 1$ means Q *redistributes* mass without a global rescale; the structured part (selection reinforces recombination biases, with the observed $P_{\text{data}}(Q)/P_{\text{gen}}(Q)$ saturating around ≈ 7 [6]) reshapes the ball mass per sequence. Separately, the *physical* thymic acceptance fraction—the fraction of recombined cells that survive to the naive repertoire—is $\alpha \lesssim 15\%$ (consistent with 10–30% for positive selection and $\approx 5\%$ for full selection) [6], and selection cuts repertoire diversity by ≈ 6 bits ($\sim$ 50-fold). Two consequences for the E-value. (i) The empirical control P_0 already *is* the post-selection law of (6), so the control estimator $\hat{\pi}$ needs no Q and no α; Q enters only the analytic estimator, where one uses $Q\, P_{\text{gen}}$ in place of P_{gen}. (ii) The global acceptance fraction α is sequence-independent and *cancels* in every ratio and in $\hat{\pi}$ (which calibrates against the control's own size M); it would matter only for an *absolute* naive-frequency estimate $f(\sigma) = \alpha\, Q(\sigma) P_{\text{gen}}(\sigma)$, e.g. when the $\hat{\pi}_{\text{gen}}$ fallback for a rare query (§7) is read as an expected count of cells rather than a probability.

Lemma 3 (The two nulls differ)**.** *In general $P_0 \neq P_{\text{gen}}$: thymic and peripheral selection deplete some motifs while finite-sample public-clone sharing enriches others, so neither $\pi_0 \leq \pi_{\text{gen}}$ nor the reverse holds universally. Hence $\hat{\pi}_{\text{gen}}$ is used as a variance-reducing control variate and as a fallback for queries too rare for the control (§7), not as a substitute for $\hat{\pi}$.*

4 Poisson approximation with an explicit error bound

Fix q, θ. For the unique clonotypes $x_1, \ldots, x_N$ of D set $X_i = \mathbf{1}(x_i \in B_\theta(q))$, $p_i = \mathbb{E}X_i = \pi_0$, $W = \sum_i X_i$, $\lambda = \sum_i p_i = N\pi_0$, and let $Z \sim \text{Poisson}(\lambda)$. We use the following standard objects. $\mathcal{L}(W)$ denotes the *law* (probability distribution) of W. The *total-variation distance* between two laws μ, ν on $\mathbb{Z}_{\geq 0}$ is

$$\mathrm{d_{TV}}(\mu, \nu) = \sup_{A \subseteq \mathbb{Z}_{\geq 0}} |\mu(A) - \nu(A)| = \tfrac{1}{2} \sum_{k \geq 0} |\mu(k) - \nu(k)|, \tag{7}$$

so a bound on $\mathrm{d_{TV}}$ bounds the error of *every* event probability simultaneously. A family of *dependency neighbourhoods* is a choice, for each index i, of a set $B_i \ni i$ such that X_i is independent of (or nearly independent of) $\{X_j : j \notin B_i\}$; intuitively B_i collects the clonotypes whose ball-membership is statistically coupled to x_i's (here, those sharing a motif). The residual b_3 below measures exactly how far that near-independence falls short.

Theorem 1 (Chen–Stein bound [3, 4])**.** *For any dependency neighbourhoods $\{B_i \ni i\}$,*

$$\mathrm{d_{TV}}\big(\mathcal{L}(W), \text{Poisson}(\lambda)\big) \leq b_1 + b_2 + b_3, \tag{8}$$

with $b_1 = \sum_i \sum_{j \in B_i} p_i p_j$, $b_2 = \sum_i \sum_{j \in B_i, j \neq i} \mathbb{E}[X_i X_j]$, and $b_3 = \sum_i \mathbb{E}\big|\, \mathbb{E}[X_i - p_i \mid \sigma(X_j : j \notin B_i)]\, \big|$.

Corollary 1 (Le Cam regime [9])**.** *If the collapsed clonotypes are independent under H_0, take $B_i = \{i\}$; then $b_2 = b_3 = 0$ and*

$$\mathrm{d_{TV}}\big(\mathcal{L}(W), \text{Poisson}(\lambda)\big) \leq \sum_i p_i^2 = N\pi_0^2 = \lambda\, \pi_0. \tag{9}$$

The bound is small precisely in the regime of interest: a rare ball $\pi_0 \ll 1$ with moderate λ gives error $\leq \lambda\pi_0 \to 0$.

Corollary 2 (Void and tail probabilities)**.** *With $w = n_D(q,\theta)$ observed,*

$$p_{\text{any}}(q,\theta) = \mathbb{P}(W \geq 1) = 1 - e^{-\lambda} + O(N\pi_0^2), \tag{10}$$

$$p(q,\theta) = \mathbb{P}(Z \geq w) = 1 - \sum_{k<w} \frac{e^{-\lambda}\lambda^k}{k!}, \quad |\,\mathbb{P}(W \geq w) - p(q,\theta)\,| \leq b_1 + b_2 + b_3. \tag{11}$$

Both follow from Theorem 1: the void probability is the event $A = \{0\}$ and the tail is $A = \{w, w+1, \dots\}$, and by (7) the error on any single event is at most $\mathrm{d_{TV}} \leq b_1 + b_2 + b_3$ (so $O(N\pi_0^2)$ in the independent regime, Corollary 1).

Remark 3 (Where biology enters)*.* Convergent recombination makes the dependency neighbourhood B_i nontrivial: $j \in B_i$ when x_i and x_j share a high-P_{gen} motif. The term $b_2 = \sum_i \sum_{j \in B_i} \mathbb{E}[X_i X_j]$ is the excess *pairwise* ball co-occupancy. Under H_0 it is controlled by the pairwise ball mass, estimable from the control by counting *pairs* of control sequences both in $B_\theta(q)$; the Poisson regime holds when this estimate is $\ll \lambda$. The very same b_2 is inflated under H_1 (antigen-driven clusters co-occupy the ball), so it is simultaneously the null error term and the quantity carrying the signal.

5 Clonal redundancy and over-dispersion

Proposition 1 (Collapsing restores Poisson)**.** *Let the raw target carry clonotypes with multiplicities (clone sizes) m_x. Counting reads/cells in the ball gives a compound-Poisson total $T = \sum_{k=1}^{K} m_k$ with $K \sim \text{Poisson}(\lambda)$ and m_k i.i.d. $\sim G$, so $\mathbb{E}T = \lambda\mu_G$ and $\text{Var}\,T = \lambda\,\mathbb{E}[m^2]$, with over-dispersion index $\text{Var}\,T/\mathbb{E}T = \mathbb{E}[m^2]/\mu_G \geq 1$. Collapsing to unique clonotypes is the projection $G \equiv 1$, which removes multiplicity-driven over-dispersion and returns the Poisson count W of §4. We therefore deduplicate C and D to unique clonotypes by default.*

Proposition 2 (Negative-binomial robustness check)**.** *If multiplicities must be modelled (e.g. read-level tests) and G is geometric, T is negative-binomial; report the NB tail $\mathbb{P}(\text{NB} \geq w)$ with mean $\lambda\mu_G$ and dispersion estimated from observed clone sizes. Under H_1 antigen-driven clones are stochastically larger, so power is retained; the collapsed Poisson test remains the assumption-light default.*

Proposition 3 (tf–idf is self-information weighting)**.** *Weighting each target hit x by its background self-information $w(x) = -\log P_0(\{x\}\text{-ball})$ makes the expected per-hit contribution constant under H_0; the inverse-document-frequency weight is exactly the inverse background ball mass, and the term frequency is the clone multiplicity. In the rare regime the control-set E-value satisfies $E \approx e^{-\sum_x \text{idf}(x)}$, so the "control-set" and "tf–idf" approaches to redundancy are one object.*

6 The E-value and multiple testing

Definition 2 (E-value)**.** For a query family $\mathcal{Q}$, the expected number of background hits is

$$E_{\text{tot}}(\theta) = \mathbb{E}_{H_0}[\#\text{hits}] = \sum_{q \in \mathcal{Q}} N\,\pi_0(q,\theta) = \sum_{q \in \mathcal{Q}} \lambda(q,\theta). \tag{12}$$

The per-query specialization $\mathcal{Q} = \{q\}$ gives the BLAST-convention E-value $E(q,\theta) = N\pi_0(q,\theta)$, estimated by

$$\widehat{E}(q,\theta) = \frac{N}{M}\, n_C(q,\theta), \qquad p_{\text{any}} = 1 - e^{-\widehat{E}}. \tag{13}$$

Proposition 4 (Assumption-free expectation)**.** *Equation* (12) *holds by linearity of expectation regardless of any dependence among hits (clonal, convergent, or across $\mathcal{Q}$). Consequently $\mathbb{P}(\#\textit{false hits} \geq 1) \leq E_{\text{tot}}$ by Markov's inequality, and E_{tot} bounds the expected number of false discoveries. This robustness—no independence needed for the* mean*—is why the E-value, not the Poisson tail, is the primary report.*

Proposition 5 (Family-wise and false-discovery control)**.** *Two thresholding regimes for a family of $|\mathcal{Q}|$ tested queries:*

1. E-value / Bonferroni (FWER). *Reporting every query with $\widehat{E}(q,\theta) \leq \alpha/|\mathcal{Q}|$ controls the family-wise error rate at level α: by Proposition 4 the expected number of false positives is $\sum_q \widehat{E} \leq \alpha$, and $\mathbb{P}(\geq 1$ false positive$) \leq \alpha$ by Markov. No independence is required. A fixed E-value cutoff (e.g. $\widehat{E} \leq 1$, the BLAST default) is the $\alpha = |\mathcal{Q}|$ case and bounds the* expected count *of false positives by* 1.

2. p-value / Benjamini–Hochberg (FDR). *Using the per-query enrichment p-values $p(q,\theta) = \mathbb{P}(Z \geq n_D^{>}(q,\theta))$ from Corollary 2, the Benjamini–Hochberg procedure [5]—sort $p_{(1)} \leq \cdots \leq p_{(|\mathcal{Q}|)}$, reject the k largest with $p_{(k)} \leq \frac{k}{|\mathcal{Q}|}\alpha$—controls the false discovery rate at α under positive dependence of the test statistics, the relevant regime here (convergent clusters induce positive correlation).*

Proposition 6 (Detectability / minimum cluster size)**.** *Under $H_1(q)$ let the antigen-driven excess add k neighbours beyond the background mean $\lambda = \lambda^{>}(q,\theta)$, so $n_D^{>} \approx \lambda + k$. The enrichment test at E-value cutoff $\widehat{E} \leq \alpha$ rejects when $n_D^{>} \geq w_\alpha$, the smallest w with $\mathbb{P}(\text{Poisson}(\lambda) \geq w) \leq \alpha$. For small λ (the typical rare-ball regime), w_α grows only logarithmically, $w_\alpha \approx \dfrac{\log(1/\alpha)}{\log\log(1/\alpha) - \log\lambda}$ by the Poisson right tail, so a cluster of a handful of convergent neighbours is already detectable; for moderate λ the Gaussian approximation gives the familiar $k \gtrsim z_{1-\alpha}\sqrt{\lambda}$. The control size enters only through the resolution of $\widehat{\lambda}$ (§7): M must be large enough that the sampling noise of $\widehat{E}$ is below the excess k being claimed.*

6.1 Epitope detection complexity

Proposition 6 concerns one query; in practice one samples a depth-n repertoire and asks how much of an epitope-specific response is recoverable. Let an epitope's TCR repertoire R_e have K unique clonotypes and within-set scope-θ neighbour graph with degree distribution $\{d_x\}_{x \in R_e}$ and neighbour density $\rho = \frac{1}{K(K-1)}\sum_x d_x = \overline{d}/(K-1)$ (the probability that two random members of R_e are within θ).

Proposition 7 (Detection curve from the degree law)**.** *Draw n clonotypes i.i.d. from R_e. A node of full-set degree d_x retains in expectation $d_x\,(n-1)/(K-1)$ of its neighbours (hypergeometric sampling). Against the near-empty background ball ($\lambda \approx 0$, so w_α is $O(1)$, Proposition 6), the node is detected once this exceeds a level $d_{\min}(\alpha) = O(1)$, i.e. at sampling depth*

$$n_x^* \approx 1 + d_{\min}\,\frac{K-1}{d_x}, \tag{14}$$

and the detectable fraction at depth n is

$$\varphi(n) = \frac{1}{K}\,\#\Big\{x : d_x \geq d_{\min}\tfrac{K-1}{n-1}\Big\}, \tag{15}$$

fixed entirely by the degree law. Equivalently the expected number of within-sample neighbour pairs is $\binom{n}{2}\rho$*, so the first significant pair appears near* $n \approx \sqrt{2/\rho}$*. The* detection complexity *of* R_e*—the depth to recover a target fraction of the response—is therefore set by the upper tail of* $\{d_x\}$ *(equivalently by* ρ *and the largest cluster): a repertoire dominated by one large convergent cluster is detected at small* n*, a diverse repertoire of many near-singletons requires deep sampling.*

Remark 4 (Worked example: A*02 NLV vs GIL). Measured on VDJdb TRB / HLA-A*02 repertoires against a 10^6-sequence OLGA background at scope $\theta = 1$ substitution: **GIL** (GILGFVFTL, influenza M1; $K = 5236$) has $\rho = 3.4 \times 10^{-4}$, max degree 52, and one dominant component of 896 (17% of the set); **NLV** (NLVPMVATV, CMV pp65; $K = 13044$) has $\rho = 2.8 \times 10^{-5}$ ($\approx 12\times$ sparser), max degree 22, and a largest component of only 152 (1.2%). Equation (15) then predicts—and the subsampled Benjamini–Hochberg significant fraction confirms (Fig. 1, `bench/bench_epitope.py`)—that GIL is ∼20–30% recovered by $n \sim 10^3$ sampled TCRs while NLV stays below 5% even at $n \sim 5 \times 10^3$. The two epitopes have detection complexities differing by more than an order of magnitude purely from repertoire structure, with no change to the search or the background.

7 How large must the control be?

Since $M\hat{\pi} \sim \text{Binomial}(M, \pi_0)$, $\text{Var}\,\hat{\pi} = \pi_0(1-\pi_0)/M$ and the relative error is $\text{CV}(\hat{\pi}) \approx (M\pi_0)^{-1/2}$.

Proposition 8 (Resolution)**.** *To resolve a target E-value* $E^* = N\pi_0$ *to relative error* ρ *requires*

$$M \gtrsim \frac{1}{\rho^2 \pi_0} = \frac{N}{\rho^2 E^*}. \tag{16}$$

Resolving $E^* \sim 1$ *to* 10% *thus needs* $M \sim 100\,N$*.*

Proposition 9 (Empty-ball regime)**.** *If* $n_C = 0$*, the point estimate* $\hat{\pi} = 0$ *is degenerate; use the rule of three* $\pi_0 \lesssim 3/M$ *(95%), or a* $\text{Beta}(n_C + a,\, M - n_C + b)$ *posterior, which propagates control uncertainty into a Poisson–Gamma (negative-binomial) posterior-predictive p-value for* n_D*. The implementation reports the rule-of-three upper bound* $\widehat{E} \le 3N/M$ *when* $n_C = 0$*.*

When M is inadequate for a rare q, the analytic $\hat{\pi}_{\text{gen}}$ is exact per query for any M and serves as the fallback (Lemma 3).

8 Composition and length are handled automatically

Proposition 10. *Because* $\pi_0(q, \theta) = P_0(B_\theta(q))$ *is computed for the* specific *query* q*, the control estimator* $n_C(q, \theta)/M$ *conditions on* q*'s length and composition automatically: the same biases that make* q *common make* n_C *large. This is the finite-sample, composition-exact analogue of the Karlin–Altschul* $K\,mn$ *length normalization, which is needed precisely because the i.i.d. background is query-independent. The only caveat is statistical: rare* q *require adequate* M *(§7), else fall back to* $\hat{\pi}_{\text{gen}}$*.*

9 The closest hit: an extreme-value law

Theorem 2 (Poisson ⇒ Gumbel)**.** *Let* $\lambda(q, t) = N\,P_0(B_t(q))$*. By the Poisson approximation applied at each radius,*

$$\mathbb{P}\big(S_{\min}(q) > t\big) \approx e^{-\lambda(q,t)} = e^{-NP_0(B_t(q))}. \tag{17}$$

If $\log P_0(B_t(q)) \approx a + \beta t$ *(log-linear ball-mass growth, the generic regime), then the best score* $Y = -S_{\min}$*, centred at* $u_N = (\log N + a)/\beta$*, obeys* $\mathbb{P}(Y - u_N \le y) \to \exp(-e^{-\beta y})$*, a Gumbel law with scale* $1/\beta$*. Here* β *is the* empirical *ball-mass log-slope (regress* $\log n_C(q,t)$ *on* t*), not the Karlin–Altschul* λ^**. (For lattice scores the Gumbel carries the usual periodic correction.)*

10 Relation to Karlin–Altschul

Theorem 3 (KA is the product-measure, ungapped case)**.** *If* $P_0 = \bigotimes_\ell p$ *is a product measure and* s *is the ungapped additive score, then* $P_0(B_\theta(q))$ *factorizes and, by Cramér's theorem,* $-\frac{1}{|q|}\log P_0(B_\theta(q)) \to \lambda^*$*, the Karlin–Altschul parameter solving* $\sum_{ij} p_i p_j e^{\lambda^* s_{ij}} = 1$*. The intensity* $\lambda(q,\theta) = N P_0(B_\theta(q))$ *then reduces to* $E = K\,m\,n\,e^{-\lambda^* S}$ *with* K *the Poisson-clumping constant, recovering [7, 8, 2].*

Thus the present framework generalizes Karlin–Altschul in three ways: (i) the product measure $\bigotimes p$ is replaced by the empirical/generative background P_0; (ii) gaps and matrix-weighted balls are admitted via the engine's score; (iii) the asymptotic constants K, λ^* are replaced by a finite-N, finite-M non-asymptotic error bound (Theorem 1).

11 The epitope case: a limitation

Remark 5. For TCR CDR3, P_0 is generation/repertoire-driven and a healthy-donor control instantiates it. For *epitopes* (MHC-presented peptides) the relevant background is presentation, not V(D)J generation. The machinery of §§2–9 applies verbatim with $P_0 := P_0^{\text{pep}}$ and a presented-peptide control, but: (L1) there is no closed-form V(D)J-style generation model P_{gen}, so only the empirical estimator survives; (L2) presentation is HLA-restricted, so P_0^{pep} is allele-conditional and the control must be HLA-matched or marginalized over an HLA frequency distribution; (L3) anchor-residue structure argues for position-weighted ball geometry. We therefore claim soundness for epitopes only to the extent that a faithful presented-peptide control is available; no generation-based null is claimed there.

12 Practical defaults and algorithm

1. Deduplicate C and D to unique clonotypes (Proposition 1).
2. Build a `seqtree` index of C; for each query compute n_C and n_D at scope θ via batched search. When the query may itself be in D or C, use the punctured counts $n^{>}$ that drop distance-zero (exact/self) hits (Lemma 2).
3. Report $\widehat{E} = (N/M)\, n_C^{>}$ (Eq. (13)), $p_{\text{any}} = 1 - e^{-\widehat{E}}$, and $p_{\text{enrich}} = \mathbb{P}(\text{Poisson}(\widehat{E}) \ge n_D^{>})$; use the rule of three when $n_C^{>} = 0$ (Proposition 9).
4. Across a query family, threshold on $\widehat{E}$ for FWER control or apply Benjamini–Hochberg to the p_{enrich} for FDR control (Proposition 5).
5. Validate the Poisson regime via the pairwise co-occupancy estimate of b_2; if inflated, use the negative-binomial check (Proposition 2).
6. Size the control by Eq. (16); fall back to the model-based $\hat{\pi}_{\text{gen}} = \sum_{B_\theta(q)} q\, P_{\text{gen}}$ (Murugan model, thymic factor $q \approx 1/2.7$) for rare queries.

This is implemented in `seqtree.evalues` (with `exclude_exact` for the punctured counts), a thin layer over batched search; the control loader `seqtree.load_control` supplies a deduplicated background.

References

[1] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.

[2] Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, 1997.

[3] Richard Arratia, Larry Goldstein, and Louis Gordon. Two moments suffice for Poisson approximations: the Chen-Stein method. *The Annals of Probability*, 17(1):9–25, 1989.

[4] Richard Arratia, Larry Goldstein, and Louis Gordon. Poisson approximation and the Chen-Stein method. *Statistical Science*, 5(4):403–434, 1990.

[5] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

[6] Yuval Elhanati, Anand Murugan, Curtis G. Callan, Thierry Mora, and Aleksandra M. Walczak. Quantifying selection in immune receptor repertoires. *Proceedings of the National Academy of Sciences USA*, 111(27):9875–9880, 2014.

[7] Samuel Karlin and Stephen F. Altschul. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proceedings of the National Academy of Sciences USA*, 87(6):2264–2268, 1990.

[8] Samuel Karlin and Stephen F. Altschul. Applications and statistics for multiple high-scoring segments in molecular sequences. *Proceedings of the National Academy of Sciences USA*, 90(12):5873–5877, 1993.

[9] Lucien Le Cam. An approximation theorem for the Poisson binomial distribution. *Pacific Journal of Mathematics*, 10(4):1181–1197, 1960.

[10] Anand Murugan, Thierry Mora, Aleksandra M. Walczak, and Curtis G. Callan. Statistical inference of the generation probability of T-cell receptors from sequence repertoires. *Proceedings of the National Academy of Sciences USA*, 109(40):16161–16166, 2012.

[11] Mikhail V. Pogorelyy and Mikhail Shugay. A framework for annotation of antigen specificities in high-throughput T-cell repertoire sequencing studies. *Frontiers in Immunology*, 10:2159, 2019.

[12] Paul-Gydeon Ritvo, Ahmed Saadawi, Pierre Barennes, Valentin Quiniou, Wahiba Chaara, Karim El Soufi, Benjamin Bonnet, Adrien Six, Mikhail Shugay, Encarnita Mariotti-Ferrandiz, and David Klatzmann. High-resolution repertoire analysis reveals a major bystander activation of Tfh and Tfr cells. *Proceedings of the National Academy of Sciences USA*, 115(38):9604–9609, 2018.

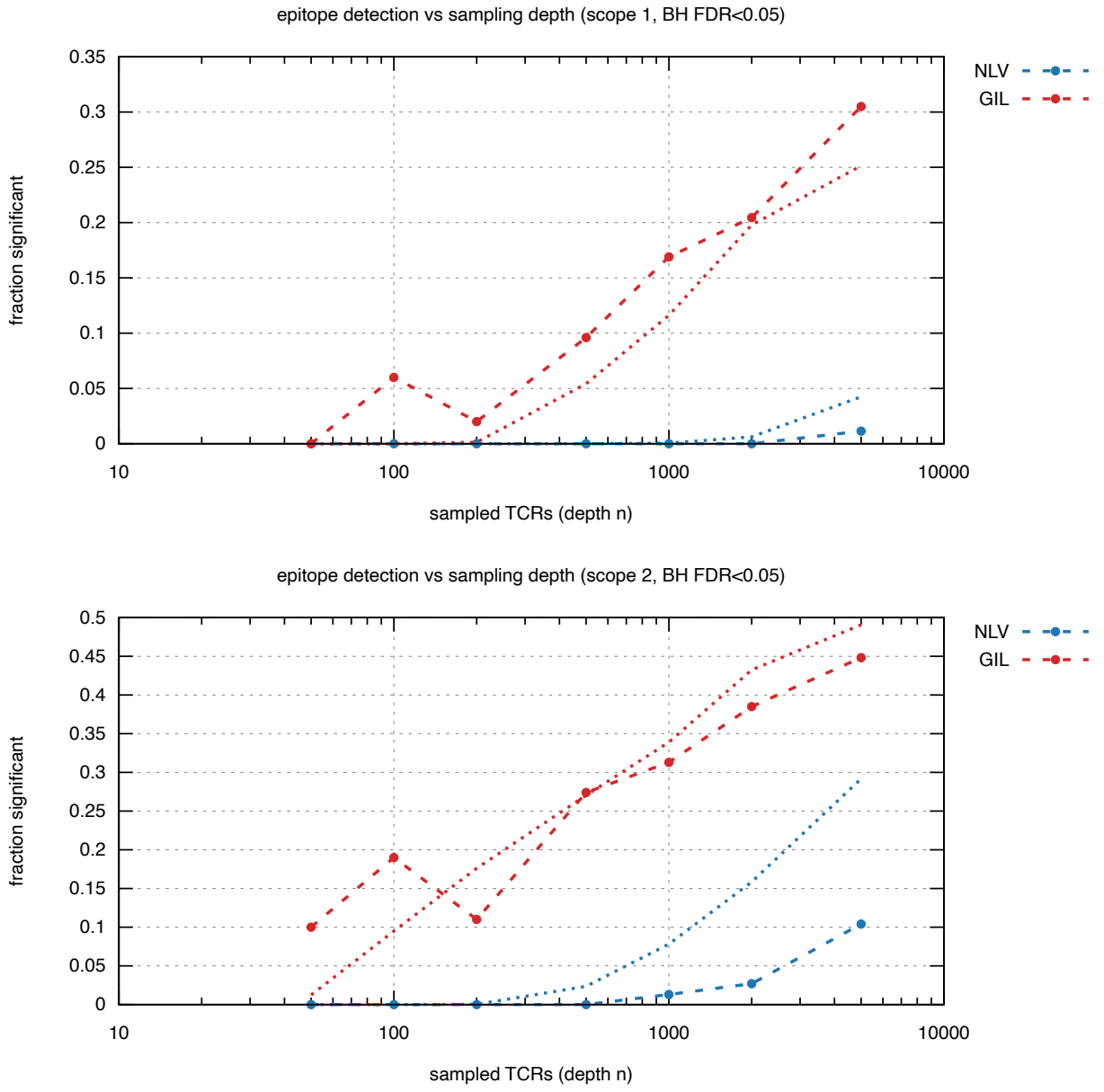

Figure 1: Epitope detection complexity (`bench/bench_epitope.py`): fraction called significant (Benjamini–Hochberg, FDR < 0.05) versus sampled depth n for the convergent GIL and the diffuse NLV A*02 repertoires. **Dashed** = observed; **dotted** = the degree-distribution prediction of Eq. (15); colour denotes the epitope (NLV / GIL). GIL's dominant cluster is recovered an order of magnitude sooner than NLV's diffuse one.