

Certified likelihoods for the distribution of fitness effects: exact site-frequency spectra with exact gradients, and error-controlled spectra under dominance

Matthew D. Schwartz
Department of Physics, Harvard University

July 9, 2026

Abstract

Inference of the distribution of fitness effects (DFE) rests on the Poisson Random Field expected site-frequency spectrum, evaluated inside likelihood loops by *POLYDFE*, *FITdadi*, and *FASTDFE* in double precision, with no error control and no analytic derivatives. We show the difficulty is structural: the three-term (contiguity) recurrence behind the spectrum carries a companion solution whose interior magnitude reaches 10^{304} at $(n, S) = (500, -100)$, so no fixed-precision scheme can evaluate the whole spectrum vector reliably at any sample size of modern interest. We then solve the problem exactly: the full vector is an element of $\mathbb{Q}(e^{-S})^{n-1}$, computable by two rational tridiagonal solves, with every S -derivative one further solve; interval (“ball”) arithmetic extends certified evaluation with gradients to $n = 10^5$, and a certified kernel cache makes DFE likelihoods as fast as the incumbents’ cached kernels. Refitting synthetic and published examples end to end: *POLYDFE* positive-selection estimates are corrupted when appreciable beneficial mass lies above $S \approx 38$ (and its shipped binary exits at every even $n \geq 90$ we tested); *FITdadi*’s 2017 tail rule carries a +74% asymptotic scale bias on its own example’s DFE. In the other direction, the modern defaults of *FITdadi* and *FASTDFE*, and *dadi*’s diffusion layer, are parameter-level sound at current data sizes. Finally, the spectrum with arbitrary dominance — previously without closed form — is an entire (everywhere-convergent), rank-three deformation of the semidominant Kummer line, giving the first error-controlled dominance spectra. We certify the equilibrium selection and mixture layers these likelihoods share; demographic models upstream are not certified here.

Contents

1	Introduction	2
2	The exact structure of the selection SFS	5
2.1	Closed form and conventions	5
2.2	Contiguity and the conditioning theorem	5
2.3	The exact solution over $\mathbb{Q}(e^{-S})$	7
2.4	A catalogue of precision traps	7
3	The certified engine	8

4	Audit of the production stack	9
4.1	POLYDFE	9
4.2	FIT <i>dadi</i> / <i>dadi</i>	9
4.3	FASTDFE	10
5	Do the failures move inferences? End-to-end refits	10
6	Selection with dominance	13
7	Discussion	15

1 Introduction

The distribution of fitness effects is a central quantity in population genetics: it enters estimates of the deleterious mutation load, the rate of adaptive evolution, the interpretation of genome-wide association data, and essentially every quantitative statement about how selection shapes standing variation. The workhorse framework for inferring it from polymorphism data is the Poisson Random Field (PRF) of Sawyer and Hartl (1992) [1], in which the expected number of sites where a derived allele appears i times in a sample of n chromosomes, at scaled selection coefficient $S = 4N_e s$, is

$$m_i(S) = \theta C(n, i) \int_0^1 x^{i-1} (1-x)^{n-i-1} \frac{1 - e^{-S(1-x)}}{1 - e^{-S}} dx. \quad (1)$$

Here $C(n, i)$ denotes the binomial coefficient $\binom{n}{i}$. A DFE model $\phi_\omega(S)$ is then mixed over S , and the parameters ω (a Gamma shape and mean for deleterious mutations, a weight and scale for beneficial ones, and nuisance parameters) are estimated by maximizing a Poisson likelihood over the observed spectrum. This is the computational core of POLYDFE [2], of FIT*dadi* [4] built on *dadi* [5], and of FASTDFE [6].

The integral in Eq. (1) has a closed form,

$$m_i(S) = \theta \frac{n}{i(n-i)} \frac{1 - {}_1F_1(n-i; n; -S)}{1 - e^{-S}}, \quad (2)$$

with ${}_1F_1$ the confluent hypergeometric (Kummer) function. We make no claim to this identity: it appears as Eq. (24) of Živković, Steinrücken, Song, and Stephan (2015) [7], and was derived independently in the supplementary material of POLYDFE [3], in both cases in passing and in neither case with any discussion of numerical error. It is implemented in production in POLYDFE (through the GNU Scientific Library’s (GSL) [22] integer-parameter Kummer routine, in double precision) and in FASTDFE (through `mpmath` at default precision, cast to double and tabulated on a fixed grid). What we have been unable to find anywhere in this literature — in a sweep of the 1992–2026 methods literature and an audit of the production codebases themselves, with one closed-access exception scoped in Section 7 — and what this paper supplies, is the following: 1) evaluation of Eq. (2), and of whole SFS vectors, with *certified* error bounds at every sample size and selection strength of applied interest; 2) *exact* derivatives of the likelihood with respect to S and the DFE parameters (every production tool uses finite differences or derivative-free search; POLYDFE’s own supplement states that “first and second derivatives [are] calculated numerically” [3]); 3) a stable $\mathcal{O}(n)$ evaluation of the full vector $(m_1, \dots, m_{n-1})$ with an understanding of *why* the naive

recurrences fail; and 4) an error-controlled treatment of the DFE mixture, including its tails. Selection with dominance $h \neq \frac{1}{2}$ is in even worse shape: no finite closed form in classical special functions was known, and no error-controlled evaluation of any kind. *dadi* reaches $h \neq \frac{1}{2}$ through its diffusion solver (the route behind dominance-DFE inference in Huber et al. 2018 [19]); the one tool that evaluates the equilibrium PRF dominance integral directly (FASTDFE, since version 1.2.0) does so by nested trapezoid rules on fixed grids, in double precision, with no error estimate. We treat the dominance spectrum in Section 6 and show it is an entire (everywhere-convergent) one-parameter deformation of the Kummer line Eq. (2).

This paper certifies the equilibrium selection layer shared by POLYDFE, FIT*dadi*, and FASTDFE; the wider DFE toolbox splits cleanly by whether a tool evaluates the same object. GRAPES [14] and ANAVAR [15] work from the same Sawyer–Hartl equilibrium expectation Eq. (1), evaluated in conventional double-precision numerics, so the Kummer-line findings here — the cancellation windows, the intrinsic instability of whole-vector recurrences, and the certified kernels that replace both — bear on their selection layers directly, though we do not refit those tools. DFE-ALPHA [13] instead propagates discrete Wright–Fisher transition matrices, and MOMENTS [16] integrates a closed system of ordinary differential equations for the SFS entries; neither evaluates the Kummer line, and their numerics are out of scope. The likelihood itself is older than any of the tools: Bustamante, Wakeley, Sawyer, and Hartl [10] introduced the PRF site-frequency-spectrum likelihood we certify, and Boyko et al. [12] set the demography-plus-DFE inference pattern the modern pipelines follow. Dominance shows the same split: Williamson, Fledel-Alon, and Bustamante [11] were already fitting PRF models with arbitrary dominance by quadrature in 2004, and joint inference of dominance with the DFE is a live production target [19]; among current tools only FASTDFE evaluates the equilibrium dominance integral directly, with *dadi* reaching it through the diffusion solver, so the dominance results here have no error-controlled baseline to compare against — Section 6 provides the first.

Whether any of this matters at the sample sizes and selection strengths of applied work is a quantitative question, and a large part of this paper is devoted to answering it in both directions. On the reassuring side: we find that FIT*dadi*’s 500-node selection grid and FASTDFE’s default 2000-bin kernel are accurate to 10^{-4} – 10^{-5} per SFS entry for human-like DFEs — errors that would require $\sim 10^7$ observed counts per entry to matter, far beyond any current dataset. Those default pipelines are, for point estimation on deleterious human-style DFEs, numerically sound. On the other side, we find failures that are not matters of degree but of kind, all inside parameter ranges the tools advertise:

- POLYDFE’s deleterious-DFE likelihood must evaluate Eq. (2) in a window $|S| \sim [1.0, 5.0] \times 10^{-5}$ where the GSL Kummer routine it pins suffers total cancellation, and the window is hit hardest at the symmetric entry $i = n/2$ — which exists only at even n , hence the parity dependence. At every even n we tested from 90 to 100 — and the mechanism applies at any larger even n — the adaptive quadrature lands in this window and the shipped binary exits with an error. We reproduced the crash byte-for-byte with the authors’ own distributed executable, at the exact node reported by a user in 2019 (GitHub issue #3); POLYDFE’s bundled simulator cannot even generate an $n = 100$ dataset. Odd n escapes by parity accident. At positive S the same window corrupts silently: the mixing integrand is wrong by factors of 10^2 – 10^3 while the quadrature reports success.
- POLYDFE’s beneficial expectations at $S \gtrsim 38$ — inside the $S_b \in [0, 100]$ range its distributed configuration accepts — are silently wrong by amounts that grow with sample size, reaching

a relative error of 22.8 (off by more than a factor of twenty) at $n = 1000$. Whether this moves a fit depends on how much beneficial mass the true DFE puts above $S \approx 38$; Section 5 maps the boundary. Estimates of the beneficial fraction and of α built on them inherit this.

- *FITdadi*’s two shipped generations differ in how they treat DFE mass beyond the cached selection grid (drop it, or clamp it to the strongest cached spectrum). On *FITdadi*’s own distributed example ($n = 250$), the two rules — identical data, identical model — give DFE scale estimates differing by 8.6% and $E[s]$ by 7.2%, with the truth in between; the log-likelihood shift between rules is of the order of a one-degree likelihood-ratio criterion.
- *dadi*’s diffusion solver, which supplies *FITdadi*’s per- S spectra, errs by 2–3% (median) and up to 57% at strong selection under its default grids at $n = 100$, and following the manual’s own grid advice makes this catastrophically worse.

Whether these SFS-level failures move fitted DFE parameters is a separate question, which we answer by refitting synthetic and published data end to end with both the incumbent numerics and the certified engine (Section 5): the POLYDFE channels and the 2017 tail rule move fitted parameters by many confidence intervals; the diffusion layer and the modern defaults do not.

The deeper point of this paper is structural. The contiguity recurrence connecting m_i to $m_{i\pm 1}$ — the natural route to whole-vector evaluation, written down for an adjacent spectrum by Vogl and Clemente (2012) [8] — has a second solution whose interior magnitude is far larger than the physical one: 10^{85} at $(n, |S|) = (100, 10)$ and 10^{284} at $(500, 100)$, in the operational sense made precise in Section 2.2. Vogl and Clemente reported their forward recursion unusable in double precision already at $\gamma = 0.1$, $n = 10$, and switched direction heuristically; we show the instability is intrinsic — it is the condition number of the underlying two-point boundary problem, so *no* floating-point scheme at fixed precision, in any direction, with any pivoting, can evaluate the vector reliably (we exhibit backward-stable solves (the best a floating-point algorithm can do) whose residuals — the standard check of how well the equations are satisfied — sit at rounding level while the solution carries zero correct digits). The cure is to leave floating point altogether: ${}_1F_1(n - i; n; -S)$ at integer parameters and rational S lies in $\mathbb{Q} + \mathbb{Q}e^{-S}$ (a rational number plus a rational multiple of e^{-S}), and the whole vector is the exact solution of a tridiagonal system over $\mathbb{Q}(e^{-S})$ with elementary boundary values, computable by two rational solves. The rational heights then *bound* the cancellation a priori, so evaluation at any target precision is certified by construction. Derivatives to any order solve the same tridiagonal system with new right-hand sides. This is, to our knowledge, the first exact and certified treatment of the PRF selection likelihood stack, and it is fast enough for production: the exact route covers $n \lesssim 2000$ in seconds; a ball-arithmetic route certified by interval radii covers biobank sample sizes ($n = 10^5$: 140 seconds for the full vector with gradient); and a certified kernel cache evaluates DFE likelihoods in milliseconds-to-seconds, the same class as the incumbents’ cached kernels.

The paper is organized as follows. Section 2 derives the exact structure: the closed form, the contiguity system, the conditioning theorem that explains the field’s numerical folklore, the exact rational solution, and the gradient stack. Section 3 describes the certified engine and its validation gates (four independent evaluation routes agreeing to ≥ 109 digits at hostile parameters; a held-out two-route gate at 74 digits across the S range). Section 4 is the audit of the production stack, with file-and-line receipts; Section 5 converts the audit into parameter-level statements by end-to-end refits. Section 6 develops the dominance spectrum: the entire q -deformation structure (an everywhere-convergent expansion in the dominance combination $\tilde{q} = (1 - 2h)S$), two certified

evaluators, and the first error-controlled values, against which we measure the production nested-trapezoid path. Section 7 states scope and limitations (the certified layer is the equilibrium selection integral and DFE mixture; nonequilibrium demography enters real pipelines upstream of it and is not certified here), and collects open problems.

2 The exact structure of the selection SFS

2.1 Closed form and conventions

Throughout, $S = 4N_e s$ with $S > 0$ advantageous, the SFS is unfolded with derived-allele polarization, and $\theta = 4N_e \mu$ per site; we set $\theta = 1$ except where stated. POLYDFE and FASTDFE use exactly these conventions; *dadi*’s $\gamma = 2N_A s$ corresponds to $S = 2\gamma$ (its heterozygote fitness is $1 + 2sh$). We audited these maps against the source of each tool before comparing any numbers; the dictionary, with receipts, is in the supplementary material.

The derivation of Eq. (2) is one line: substitute $u = 1 - x$ in the exponential piece of Eq. (1) and recognize Euler’s integral for the Kummer function,

$$\int_0^1 x^{i-1} (1-x)^{n-i-1} e^{-S(1-x)} dx = B(i, n-i) {}_1F_1(n-i; n; -S), \quad (3)$$

with B the Beta function; the prefactors collapse to $n/[i(n-i)]$ by $C(n, i)B(i, n-i) = n/[i(n-i)]$. Three limits provide analytic gates. As $S \rightarrow 0$, $m_i \rightarrow 1/i$, the neutral spectrum. As $S \rightarrow +\infty$, $m_i \rightarrow n/[i(n-i)]$, the uniform-flux plateau. As $S \rightarrow -\infty$, $m_i \sim \frac{n}{i(n-i)} \frac{\Gamma(n)}{\Gamma(n-i)} |S|^{-i}$: mutation–selection balance, with entry i vanishing like $|S|^{-i}$. This last regime is where fixed grids and double precision quietly die — absolute errors look tiny while relative errors, which are what a likelihood ratio feels, diverge.

2.2 Contiguity and the conditioning theorem

Write $M_i \equiv {}_1F_1(n-i; n; -S)$. Gauss’s contiguous relations (classical identities linking Kummer functions whose parameters differ by one) give a three-term recurrence in i ,

$$i M_{i+1} + (n - 2i - S) M_i - (n - i) M_{i-1} = 0, \quad M_0 = e^{-S}, \quad M_n = 1, \quad (4)$$

with *elementary values at both ends*. This looks like a gift: an $\mathcal{O}(n)$ path to the whole vector. It is a trap. The recurrence has two solutions; the physical one is minimal in a direction set by the sign of S , and the companion solution grows. Vogl and Clemente [8], evaluating the analogous recurrence for the boundary-mutation spectrum, found the forward direction “useless” in double precision at $\gamma = 0.1$, $n = 10$, and adopted a backward recursion; this fixes the direction of contamination but not its size. We measured the leak directly: even in the stable direction, the recurrence loses 23 digits over $n = 100$ steps at $|S| = 10$ (a directional-contamination count, distinct from the boundary-value condition number below).

Since both boundary values in Eq. (4) are known, the textbook remedy is Olver’s method [20]: solve the two-point boundary problem as a tridiagonal linear system. This too fails at fixed precision, and it fails in an instructive way. The condition number of the boundary-value map is set by the *interior* magnitude of the companion solution. Concretely: writing the exact solution as $M_i = \alpha_i + \beta_i e^{-S}$ with unit-normalized boundary data, the companion magnitude $\max_i |\alpha_i|$ reaches

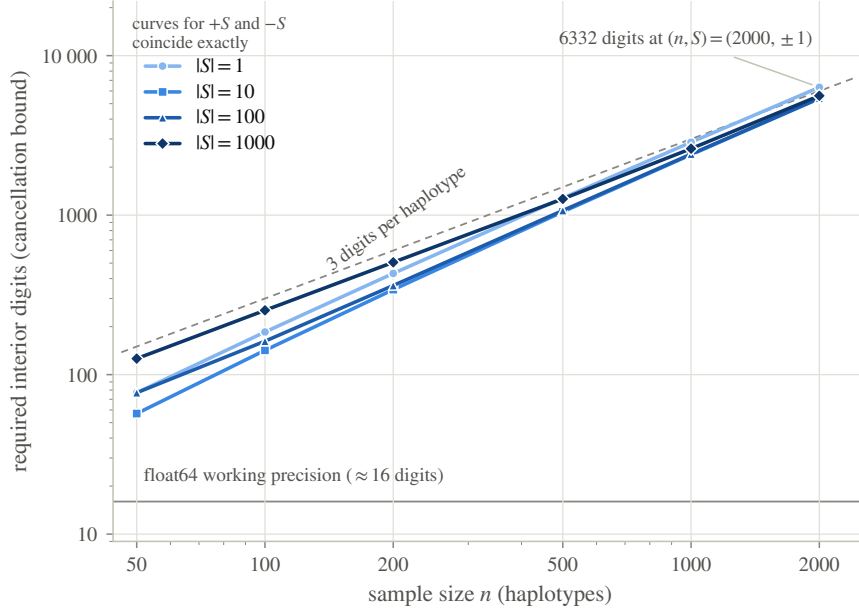


Figure 1: Why fixed precision cannot evaluate the selected SFS vector at modern sample sizes. Each curve is the interior working precision, in decimal digits, required by the two-point boundary problem for $M_i = {}_1F_1(n-i; n; -S)$: the a priori cancellation bound, measured exactly as the bit length (‘height’) of the exact rational solution over $\mathbb{Q}(e^{-S})$ times $\log_{10} 2$, on the 48-cell grid $n \in \{50, \dots, 2000\}$, $S \in \{\pm 1, \pm 10, \pm 100, \pm 1000\}$. The requirement is exactly symmetric under $S \rightarrow -S$ (all 24 pairs agree bit for bit), so one curve per $|S|$ is drawn. It grows at about three digits per haplotype (dashed guide), is largest near $|S| \sim 1$ (6332 digits at $(2000, \pm 1)$), and dwarfs the ≈ 16 digits of double precision (solid rule) already at $n = 50$. The bound is invisible to residual checks (Section 2.2); the certified engine measures it per instance and pays exactly what each instance requires.

10^{83} – 10^{88} at $(n, S) = (100, \pm 10)$ and 10^{304} at $(500, -100)$, and the operational condition number measured by a working-precision ladder (the precision below which all accuracy is lost) is 10^{85} and 10^{284} at those points, matching $\max_i |\alpha_i/M_i|$. A backward-stable pivoted solve at $(500, -100)$ returns a residual at rounding level (10^{-95} at 95-digit working precision) and zero correct digits. There is no precision at which one can simply “be careful”: the required working precision grows with n and is invisible to residual checks. We measured it two ways. The exact route of Section 2.3 computes the boundary-value problem’s cancellation bound directly (the rational heights of its solution): about three digits per sample at weak selection, e.g. 6332 digits at $(n, S) = (2000, \pm 1)$, largest near $|S| \sim 1$ and mildly $|S|$ -dependent (Figure 1; full grid in the supplement). The ball route pays the one-directional contamination, which is smaller but still enormous: 2 – 4×10^4 digits at $n = 10^5$. Both are measured per instance by the engine itself; the certified routes pay exactly what each instance requires, rather than guessing.

Two scope statements. First, the theorem is about *whole-vector* routes (recurrences and boundary-value solves) — the per-entry evaluation the production tools actually use is not subject to it, and indeed the Kummer-reflected positive-term series is stable in plain double precision for

deleterious S . Second, the production failures documented in Section 4 have their own per-entry causes (a library branch defect, float64 underflow), not this conditioning phenomenon; what the conditioning explains is why no stable $\mathcal{O}(n)$ whole-vector or gradient evaluation existed to retreat *to* — the regime that matters at biobank n — and why the one documented attempt at a recurrence (Vogl and Clemente’s) had to switch directions heuristically.

2.3 The exact solution over $\mathbb{Q}(e^{-S})$

The resolution is that Eq. (4) should not be solved in floating point at all. For rational S , every M_i lies in the two-dimensional $\mathbb{Q}$ -vector space spanned by $\{1, e^{-S}\}$:

$$M_i = \alpha_i + \beta_i e^{-S}, \quad \alpha_i, \beta_i \in \mathbb{Q}, \quad (5)$$

and the boundary values $M_0 = e^{-S}$, $M_n = 1$ split the tridiagonal system into two rational solves — one for α , one for β — performed exactly. The companion solution’s enormous interior magnitude reappears as the *heights* of the rationals α_i, β_i (their bit lengths), and this is now a feature: the heights are computed alongside the solution and *bound the cancellation in Eq. (5) a priori*, so assembling M_i at any target precision with a height-dependent working precision is certified by construction. At $n = 1000$ the two solves and the gradient solve below take 3.4 seconds.

Derivatives come from the same matrix. Differentiating Eq. (4) k times with respect to S leaves the tridiagonal operator fixed and produces the source $k M_i^{(k-1)}$ on the right-hand side, with elementary boundary values $M_0^{(k)} = (-1)^k e^{-S}$, $M_n^{(k)} = 0$. Each derivative order of the *entire* vector is therefore one additional rational tridiagonal solve. No production tool has any analytic derivative of its likelihood; here the full gradient and Hessian stack is exact, and we use it both for optimization and for observed-information confidence intervals in Section 5.

Two practical remarks. First, the exact route is comfortable to $n \approx 2000$; beyond that the height growth makes it expensive, and we switch to a ball-arithmetic recursion in the stable direction, in which interval radii track the contamination exactly and the working precision is doubled until the certified radius meets target. This is how the engine reaches $n = 10^5$. Second, the sign-dependent stable direction needs seed values one step from the boundary; these are elementary, ${}_1F_1(1; n; z) = (n-1)! z^{1-n} (e^z - \sum_{m \leq n-2} z^m / m!)$, with a series/elementary crossover at $z \approx 2n$ chosen to avoid the Poisson-tail cancellation of the elementary form (a 1008-digit cancellation at $(n, z) = (1000, 37.5)$ if used naively — the kind of trap this problem sets at every turn; we catalogue several more in the supplement).

2.4 A catalogue of precision traps

We record here, compactly, the numerical traps we hit *ourselves* while building the certified engine — partly as a user’s guide to the supplement, and partly because they are direct evidence for the paper’s thesis that this problem class cannot be trusted to uncertified numerics. 1) A general-purpose adaptive quadrature (mpmath’s, both Gauss–Legendre and tanh-sinh rules) silently underconverges on smooth exponential-ramp panels of this problem, returning 1%-wrong panel integrals with optimistic internal error estimates, frozen across its precision ladder. 2) Two evaluation routes that agree to 150 digits can both be wrong: our two Kummer forms of Eq. (2) share the reflection mpmath applies internally, so their agreement is not independence (we replaced one leg with an exact-rational certified series). 3) Any quantity evaluated at two *representations* of the same S

— a decimal and its binary-float image — differs by the full cancellation factor of Section 2.2; threading one exact rational S through every route is mandatory (a 10^{-16} representation mismatch cost 7000 digits in one of our gate runs, and a subtler instance capped a quadrature node cache at 18 digits against a 40-digit target). 4) Endpoint-singular DFE densities (Gamma shape $b < 1$) put an algebraic singularity under a Gauss rule if the neutral end of the mixing integral is handled naively. Each trap has a measured incident behind it; none survived the gates of Section 3.

3 The certified engine

The engine has three evaluation routes with disjoint failure hypotheses, and every shipped number is gated by agreement between at least two of them.

Route E (exact). The rational tridiagonal solves of Section 2.3, assembled at height-certified precision. Values and all derivative orders. $n \lesssim 2000$.

Route B (ball). Stable-direction recursion in arb ball (rigorous midpoint-plus-error-radius interval) arithmetic; the certificate is the interval radius itself, with precision doubling to target. Values and gradients, n to 10^5 . Two timings recur in this paper and differ by design: the stock precision-doubling ladder solves $n = 100003$ with gradient in ~ 125 – 140 s (it overshoots, delivering $\sim 4 \times 10^4$ digits), while a jump-started solve at the ~ 230 -digit kernel target takes $6.1 + 1.5$ s.

Route D (direct). Per-entry Kummer evaluation through positive-term series arranged by Kummer reflection [24] so that no alternating cancellation appears, with certified geometric tails; used for spot gates and as the independent anchor.

What “certified” means, route by route. The word carries three different strengths in this paper and we keep them distinct. Rigorous certificates: arb ball radii (route B; interval arithmetic, error carried by construction [21]), exact rational heights (route E; the cancellation bound is computed, not estimated), and proven geometric series tails (route D and the series anchors). Validated evidence, short of a certificate: the composite Gauss–Legendre layers (kernel cache, dominance oracle) rely on degree-pair self-checks (the same integral recomputed at two quadrature orders must agree) plus agreement with independent routes — strong evidence with a known failure mode (a shared systematic passes both degrees), which is why every quadrature layer in this paper is additionally gated against an object of the first class. Library trust: the transcendental leaves (e^{-S} and the incomplete-gamma and error functions through `mpmath` [23]) are correctly-rounded to working precision by their implementations, not by us; each such leaf is anchored by at least one exact-rational spot gate. This is the distinction we press against FASTDFE in Section 4: using arbitrary-precision software is the third level, not the first.

Validation is layered, in two stages built at different times. The reference stage: at twenty parameter points including $S = \pm 1000$ and n to 1000, the closed form, a direct certified quadrature of the defining integral Eq. (1), and the exact-rational series anchor agree to at least 109 digits (quadrature leg) and 149 digits (series leg); contiguity residuals sit at 10^{-150} — a check that gates implementation consistency with the recurrence, not conditioning-induced error, which residuals cannot see. The engine stage: routes E and B, built afterwards, were gated against that anchor at fresh held-out points (below). A held-out gate at 22 fresh points across $n \in \{23, 137, 1024, 10007\}$

and $|S| \in [10^{-4}, 2000]$, including the moderate- $|S|$ band where directional recursions leak, gives worst-case agreement of 74.3 digits for values and 74.5 for gradients (the comparison ceiling of that gate). Gradients are additionally gated against the parameter-shift identity $\partial_{S1} F_1(n-i; n; -S) = -\frac{n-i}{n} {}_1F_1(n-i+1; n+1; -S)$ and the Kummer ODE at second order.

For DFE likelihoods we precompute a *certified kernel cache*: whole-SFS vectors with gradients at fixed quadrature nodes in $\log |S|$ (composite Gauss–Legendre with a degree-pair self-check, panels bounded in integrand log-range), plus an exact small- S zone where the S -expansion of Eq. (1) has exact rational coefficients and the DFE moments are incomplete-gamma closed forms. A model-C mixture (deleterious reflected Gamma plus beneficial exponential) then evaluates as a weighted dot product against the cache — milliseconds to seconds, the same complexity class as *FITdadi*’s and *FASTDFE*’s cached kernels — but with explicit tail bounds and an exact parameter-gradient stack (the analytic s -derivative of the incomplete gamma included; we verified the full gradient against high-order finite differences at the level the finite differences support, 16 digits). Deleterious-only evaluations agree with an independent implementation of the mixing integral to 53 digits with two-scheme self-consistency at the same level.

4 Audit of the production stack

This section reports what the production tools actually compute, with file-and-line receipts (collected in the supplement), and the measured error of each layer against the certified engine. We emphasize two ground rules. First, all conventions were audited before any comparison — the single largest source of fake discrepancies in cross-tool exercises. Second, every comparison below is against certified values whose own error budget is orders of magnitude below the effects reported.

4.1 polyDFE

POLYDFE evaluates Eq. (2) through `gsl_sf_hyperg_1F1_int` in double precision, with a Taylor switch at $|S| \leq 10^{-5}$ and overflow-avoiding branches; the DFE mixture uses adaptive Gauss–Kronrod quadrature with a relative tolerance of 10^{-5} and an escalation ladder that relaxes the tolerance to 10^{-1} before giving up; derivatives are finite differences (their supplement: “first and second derivatives calculated numerically”). Where the double precision closed form is healthy, the mixing quadrature is comfortably within its budget — typical deleterious-fit errors are 10^{-7} – 10^{-6} per entry, and we found no consequential quadrature-only failure at cited deleterious parameters. The failures are the poison window and the beneficial band described in the introduction, quantified in Table 1; the GSL window failure is version-pinned in an unfortunate way, since POLYDFE’s README requires GSL 1.16 (“does not work with a newer version”), so its users cannot even trade the crash for a different bug.

4.2 fitdadi/dadi

FITdadi precomputes per- γ spectra by *dadi*’s implicit finite-difference solution of the forward Kolmogorov equation, caches them on a 500-node logarithmic γ grid, and integrates the DFE by trapezoid over the grid with heuristic treatment of the mass beyond the grid: the 2017 release drops it; the current *dadi* release clamps it to the most-deleterious cached spectrum. The grid trapezoid contributes a benign common-mode $\sim 1.7 \times 10^{-4}$ for human-like DFEs. The two tail

tool	layer	typical rel. error	worst measured	failure regime
POLYDFE	fixed- S , float64	$\leq 10^{-16}$ – 10^{-12}	1.7×10^5 (silent) / crash	$ S \in [1.08, 5.0] \times 10^{-5}$ (fixed- S grid; end-to-end crash nodes from 1.0×10^{-5}), even $n \geq 90$; $S \geq 38$ high- i
POLYDFE	DFE mixing (qag)	10^{-7} – 10^{-6}	3.2×10^3 (silent in a non-crashing build; the shipped binary exits)	transports the window bug into mixtures
FIT <i>dadi</i>	γ -grid trapezoid	$+1.7 \times 10^{-4}$	1.2×10^{-2}	singletons under steep DFEs
FIT <i>dadi</i>	tail rule (2017 vs current)	10^{-4}	1.8×10^{-2} disagreement	DFE mass beyond the grid
<i>dadi</i>	PDE + extrapolation	10^{-6} – 10^{-4}	0.57; 1.5×10^5 (manual’s grids)	strong selection, high i
FASTDFE	fixed- S + kernel	10^{-9} – 10^{-5}	2.4×10^{-3} (100-bin docs claim)	none found at defaults

Table 1: Summary of the audit (Section 4): each production layer against the certified engine, over the tools’ own advertised parameter ranges. “Silent” marks failures where the tool reports success. Full per-point tables, receipts, and the biologically cited ranges are in the supplement.

rules and the PDE layer are the consequential channels (Table 1 and Section 5). Our PDE measurements are at demographic equilibrium, deliberately: that is the one setting where the exact answer is independently known, so it isolates the solver and extrapolation layer under conditions maximally favorable to it; runs with demography solve the same discretized operator and inherit at least this error class. We do not certify, and do not test, the demographic layer itself.

4.3 fastDFE

FASTDFE tabulates Eq. (2) at 2002 bin edges through `mpmath` — arbitrary-precision software, but run at its default 15 digits, cast to double, and tabulated once; using an arbitrary-precision library without error control is not certification, and the distinction is the subject of Section 2.2. It assembles a trapezoid kernel and evaluates mixtures as matrix products; h is supported by a 21-point linear interpolation grid, and $h \neq \frac{1}{2}$ spectra come from nested trapezoids (Section 6). At its defaults the semidominant pipeline is the cleanest of the three — worst kernel-layer error 2.4×10^{-5} for the DFEs we tested — and we found no consequential defect in cited ranges. Its documentation’s claim that 100 bins per sign is “nearly identical” to the default is wrong by two orders of magnitude (2.4×10^{-3}), which begins to matter only at biobank-scale counts.

5 Do the failures move inferences? End-to-end refits

SFS-entry errors are not parameter errors: a common-mode error is absorbed by the mutation-rate nuisances, and only the component differential across frequency classes, projected on the score, moves fitted DFE parameters. We therefore refit — synthetic data generated from certified truth, and the tools’ own distributed examples — with the incumbent numerics and with the certified likelihood, and compare fitted parameters against observed-information confidence intervals as a function of data size. One scope statement before the numbers: we certify the *computation* of the standard PRF likelihood. Its modeling assumptions — independent Poisson counts across

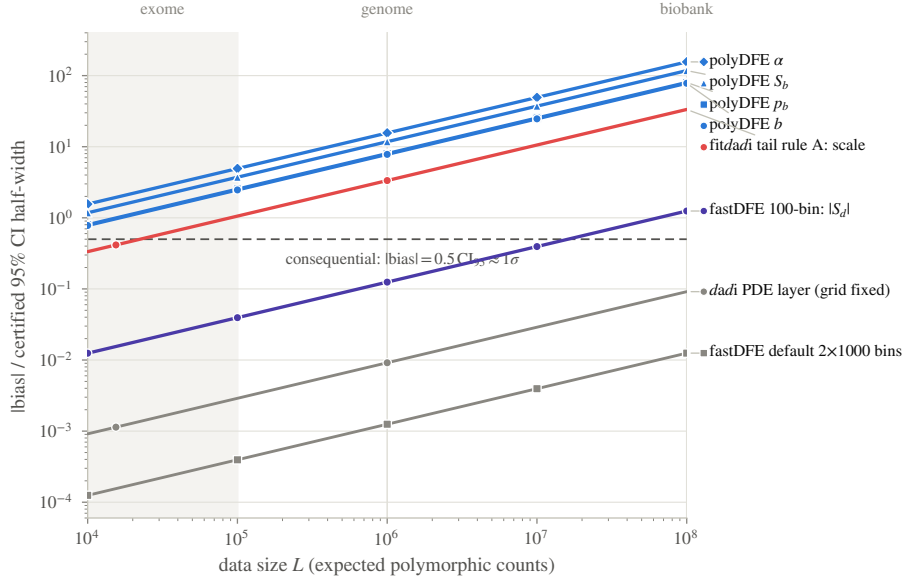


Figure 2: Fitted-parameter bias of the incumbent numerics, in units of the certified 95% confidence half-width, versus data size L (expected total polymorphic counts). Corrupted channels rise past the 1σ line (dotted) within realistic data sizes: POLYDFE’s beneficial-band channel at $n = 101$, truth $S_b = 40$ (its α_{dfe} , S_b , p_b , b curves); FITdadi’s 2017 drop rule on the shipped example’s DFE; FASTDFE at the documentation-endorsed 100-bin setting. True negatives stay flat and low: FASTDFE defaults and the *dadi* diffusion layer. Bias is a fixed offset in every channel’s construction, so curves scale as $\sqrt{L}$ between receipt points (markers); every marker traces to an archived fit.

sites, no linkage, polarization handled through the tools’ own nuisance parameters — are inherited unchanged, and nothing here bears on their adequacy; at biobank scale those assumptions, not arithmetic, may well dominate the honest error budget of any DFE fit, ours included.

The refit harness itself is validated by positive controls in every track: in POLYDFE’s healthy regime ($n = 19$, moderate deleterious selection) the incumbent and certified fits agree to 10^{-4} of a confidence interval across the entire data-size ladder; the certified side recovers its own truth to 5×10^{-6} relative or better everywhere; goodness-of-fit z -scores on the capability fits sit within $[-0.5, 1.4]$. Every number below is from a track whose control passed. Data sizes are quoted as L , the expected total polymorphic count: 10^4 – 10^5 is exome scale, 10^6 genome scale, 10^8 biobank scale. Figure 2 is the section in one picture.

polyDFE: positive-selection inference is corrupted at modern sample sizes. At $n = 101$ (odd, so no crash — nothing visible goes wrong), with a model-C truth $(b, S_d, p_b, S_b) = (0.4, -400, 0.05, 40)$, the silent Kummer collapse at $S \gtrsim 38$ drags every fitted parameter off truth while the certified refit of identical data recovers it: the shape b fits at 0.455 ($7.8\times$ the certified 95% CI at genome scale), S_d at -251 ($2.7\times$), p_b at 0.0685 ($8.0\times$), S_b at 18.5 ($11.8\times$), and α_{dfe} — the fraction of relative fixations attributable to the beneficial component, computed from the fitted DFE through the fixation kernel $S/(1 - e^{-S})$ as POLYDFE defines it — at 0.9499 against a true 0.9642: $15.6\times$ the CI. The bias equals one standard error already at $L \approx 2\text{--}4 \times 10^3$ counts, below

a single exome.

How far does this reach into published work? The corruption is a property of the beneficial mass above the $S \approx 38$ threshold, so we swept the true S_b at fixed (b, S_d, p_b) : at $S_b = 2.84$ with $p_b = 0.031$ — the values POLYDFE’s own distributed example fits — the corrupted mass is 5×10^{-8} and every parameter bias is below 10^{-3} of a CI at genome scale; at $S_b = 5$ the worst bias is 0.085 CI; at $S_b = 10$ it is $2.3 \times$ the CI, and at $S_b = 20$ the corruption reaches exome-scale data. The boundary is set by the corrupted mass $p_b e^{-38/S_b}$ crossing $\sim 10^{-3}$ (genome scale) or $\sim 7 \times 10^{-3}$ (exome scale). The honest statement is therefore scoped: fits reporting modest beneficial strength ($S_b \lesssim 5$) are safe from *this* channel, while fits reporting or exploring $S_b \gtrsim 10$ at modern n are inside it — and the even- n crash and the deleterious poison window are independent of this scoping entirely. At even $n = 100$ the authors’ binary exits during initialization on the same data (the issue-#3 signature, byte-identical), so there is no fit at all; a counterfactual build that survives the window converges to $b = 0.99$, $S_d = -46$ — a different DFE entirely, 83 times the certified 95% interval from truth in shape. And odd n does not rescue large samples: at $n = 999$ the deleterious quadrature evaluates sign-flipped double-precision residue ($\sim -10^{-288}$) at deep-deleterious nodes, 11,406 poisoned evaluations per likelihood call, on which the shipped binary again exits. The practical statement: POLYDFE model C at even $n \geq 90$ does not run at all; at odd n its positive-selection quantities are reliable only when the fitted and true beneficial mass above $S \approx 38$ is negligible, a condition the tool neither checks nor reports. Published α or p_b estimates from model C at $n \gtrsim 90$ with $S_b \gtrsim 10$ warrant rechecking; those with $S_b \lesssim 5$ do not, on this channel’s account.

fitdadi: the tail rule is the problem, not the PDE. On the shipped example’s own DFE (Gamma shape 0.186, scale 687 in γ units, $n = 250$), the 2017 drop rule carries a +74% asymptotic bias in the fitted scale and +72% in $E|s|$ — measured as the fit of noise-free certified expected data with the mutation rate profiled, the estimand the rest of this section uses; with the mutation rate held fixed at truth (the example script’s own convention) the pair is -7% (drop) and +5% (clamp). In either convention the two rules sit about one certified standard error apart at the example’s own data size and 3.4–4.0 times the CI apart at genome scale, and the drop rule’s fitted tail mass $P(|S| > 4000)$ is 1.6 standard errors from truth already at the example’s size. On a steeper DFE (mean $|S| = 10^4$) the drop rule runs away to the optimizer’s box bound. The current clamp rule is milder (-11% scale profiled, inside the CI at the example’s own size, consequential from $L \approx 5 \times 10^5$ — a number specific to the example’s narrow $|\gamma| \leq 500$ cache grid, which strands 7.5% of the DFE mass beyond its edge; the default $|\gamma| \leq 2000$ grid strands far less) but under-calls the steep DFE’s scale by 92%. The 8.6% scale split between the two shipped rules on the real example data is consistent with all of this: it is the mutation-rate-fixed, one-realization image of the same disagreement (0.6σ at the example’s size; the decomposition — 94% estimand convention, 5% demography, 2% realization — is in the supplement). Meanwhile the diffusion layer, which our SFS-entry audit flagged with errors up to 57% at strong selection, moves *no fitted parameter* by more than 0.5% once grid and tail rule are held fixed — the error is common-mode and the score projection kills it. We consider this pair of results the clearest demonstration that entry-level audits mislead in both directions and only end-to-end refits settle consequence.

fastDFE: defaults clean, documentation wrong at biobank scale. At its default 2×1000 -bin kernel the fitted-parameter bias is at most 0.03 standard errors even at $L = 10^8$ ($|\text{bias}| = 1\sigma$ would require $L \sim 10^{11}$ – 10^{12}). The documentation’s claim that 100 bins per sign is “nearly

identical” fails exactly as the kernel audit predicted: +1.6–1.8% systematic inflation of $|S_d|$, 2.4σ at $L = 10^8$ and 7.7σ at 10^9 , crossing 1σ at $L^* = 1.7 \times 10^7$. The bin-halving scaling is the expected h^2 , and the shape–scale ridge (correlation ≈ -0.9) amplifies the kernel error by a factor of about seven into the scale parameter.

Biobank sample sizes: a column the incumbents cannot fill. At $n = 50,000$ and $n = 100,000$ (with $L = 10^8$) the certified kernel builds in 4.2 and 14.7 minutes, and full maximum-likelihood refits recover all four DFE parameters within 1.2 standard errors of truth (29 and 50 minutes end to end); a single certified fixed- S SFS with gradient at $n = 10^5$ takes $6.1 + 1.5$ seconds. Two disclosures: the biobank kernels run a reduced quadrature degree pair (self-consistency ~ 8 digits — identical on the truth and fit sides, so it cancels in the bias comparison), and the optimizer stage runs on a float64 mirror of the certified kernel, re-verified against it at the optimum ($\sim 10^{-15}$). The incumbent column of this comparison is structurally empty: POLYDFE cannot run (even n ; its simulator cannot even generate the data), *dadi* requires diffusion grids finer than n with error already at the percent scale two orders of magnitude below this n , and FASTDFE’s double-precision cast floors the deep tail of the spectrum to exact zeros. Sample sizes of this order are routine in biobank exome releases; the certified engine is at present the only path to a DFE likelihood there with any error statement at all.

6 Selection with dominance

For dominance coefficient $h \neq \frac{1}{2}$ the sojourn density no longer has the form Eq. (1). With $Q(x) = px + \tilde{q}x^2$, $p = 2hS$, $\tilde{q} = (1 - 2h)S$, the expected spectrum is the two-fold integral

$$m_k(S, h) = C(n, k) \frac{\int_0^1 x^{k-1} (1-x)^{n-k-1} e^{Q(x)} \int_x^1 e^{-Q(y)} dy dx}{\int_0^1 e^{-Q(y)} dy}, \quad (6)$$

which reduces to Eq. (1) at $h = \frac{1}{2}$ ($\tilde{q} = 0$). The sojourn density here is Wright’s classical stationary formula; the sampled integral Eq. (6) is the object Williamson et al. [11] and Boyko et al. [12] evaluate numerically for PRF inference with dominance, and spectral expansions of the underlying transition density exist [17]. What the literature does not contain is a finite closed form in classical special functions or an error-controlled evaluation at arbitrary (S, h) ; among production tools, *dadi* reaches the object through its diffusion solver [19], while FASTDFE evaluates Eq. (6) directly by an outer 1000-point power-clustered trapezoid over x and an inner 200-point linear trapezoid over y , in double precision, with no error control.

The structure of Eq. (6) is that of an *entire deformation* of the Kummer line. The building block $K(a, c; p, \tilde{q}) = \int_0^1 x^{a-1} (1-x)^{c-1} e^{px + \tilde{q}x^2} dx$ satisfies the heat relation (a diffusion equation in the parameters) $\partial_{\tilde{q}} K = \partial_p^2 K$ together with contiguity in (a, c) , and the system is holonomic of rank three. The proof is one line: with $I_m \equiv \partial_p^m K$, integrating $\partial_x [x^{a+m} (1-x)^c e^{px + \tilde{q}x^2}]$ over $[0, 1]$ (the boundary terms vanish for $a+m, c \geq 1$) gives

$$(a+m) (I_m - I_{m+1}) - c I_{m+1} + p (I_{m+1} - I_{m+2}) + 2\tilde{q} (I_{m+2} - I_{m+3}) = 0, \quad (7)$$

so for $\tilde{q} \neq 0$ the span of $\{K, \partial_p K, \partial_p^2 K\}$ closes under ∂_p , and the heat relation closes the $\tilde{q}$ -direction on the same three-dimensional space (we verified Eq. (7) numerically to 10^{-60} and generic independence of the three generators). Its $\tilde{q}$ -expansion

$$K(a, c; p, \tilde{q}) = \sum_{j \geq 0} \frac{\tilde{q}^j}{j!} B(a+2j, c) {}_1F_1(a+2j; a+c+2j; p) \quad (8)$$

converges everywhere, faster than any geometric series, along the same even- a adjacent-parameter (contiguity) ladder the semidominant engine already walks. The inner integral of Eq. (6) expands in the same way with elementary coefficients, so the dominance spectrum reduces *entirely* to certified Kummer and Beta objects. In words: dominance is not a new special function; it is a one-parameter entire flow away from the object every DFE tool already computes, with the semidominant case as its initial condition.

We implemented two independent evaluators. The first, our reference evaluator (the “oracle”), evaluates Eq. (6) directly: the inner integral has a closed form by completing the square (an erf/erfi expression valid for either sign of $\tilde{q}$, which we verified against an independent certified quadrature to 86 digits), and the outer integral uses the same certified composite-Gauss–Legendre machinery as the rest of the engine. Its gates: at $h = \frac{1}{2}$ it collapses onto the exact closed form Eq. (2) to 59.9–62 digits, and its two-degree self-check holds 55–61 digits at every point of the validation grid, including $(S, h) = (-1000, 0.2)$ and $(+500, 0.7)$. The series route through Eq. (8) is the production path; its naive arrangement — elementary antiderivatives for the inner coefficients — is unstable at $|p| \lesssim 2|\tilde{q}|$ (a factorial-over-power blowup we measured at 900 lost digits before rearranging), and the stable arrangement buys uniform convergence with a computed precision pad. The two routes, disjoint in every ingredient except arithmetic — agree to 41–42 digits (the series route’s certification target at its working pad) at ten cross-gated points, spanning $S \in [-100, 500]$ and $h \in [0.05, 0.95]$ (the $(500, 0.7)$ corner at 40.7 digits with a 507-digit working pad); the remaining hostile corner $(S, h) = (-1000, 0.2)$ rests on the oracle alone, whose self-checks hold 55–61 digits there and whose $h = \frac{1}{2}$ collapse gate and closed-form inner integral are corner-independent. A third gate is convention-independent: *dadi*’s diffusion solver, handed the same (S, h) points through its own native dominance parameterization, reproduces the certified spectra at 2.3×10^{-4} relative or better — inside *dadi*’s own accuracy — including a 10^{-26} -scale tail entry tracked at 2×10^{-5} relative; an h -polarization mismatch between our convention and *dadi*’s would have appeared at up to a factor of 23, and did not. (The road there supplied one more entry for the trap catalogue: an input conversion performed outside the working-precision scope produced a uniform 16-digit ceiling between the two routes that neither route’s own two-precision self-check could see; only a third, sub-object-level arbitration localized it.)

Measured against these first error-controlled values, FASTDFE’s nested-trapezoid dominant path (1000×200 nodes, double precision) errs by 2×10^{-5} in easy regimes but degrades to 1.6×10^{-3} at $(S, h) = (-100, 0.2)$ and 2.3×10^{-2} at $(-1000, 0.2)$ — up to a thousand times worse than its own semidominant kernel, at parameter values (strong purifying selection on partially recessive mutations) where dominance-DFE inference operates. Figure 3 maps the measured error over the (S, h) receipt grid. One caution, applying our own standard from Section 5 to ourselves: these are SFS-entry errors, and we have not yet converted them into fitted- (h, DFE) bias by an end-to-end dominance refit — at strong $|S|$ the affected entries carry little likelihood weight, so the parameter-level consequence remains open. Its 21-point linear h -interpolation adds a separate, so far unmeasured, layer on top. A certified dominance kernel in the style of Section 3 removes both,

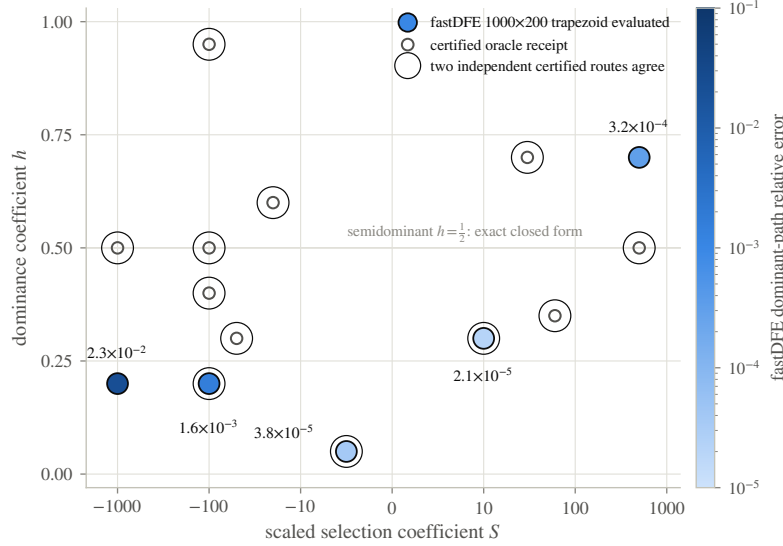


Figure 3: FASTDFE’s nested-trapezoid dominance path measured against the first error-controlled values of Eq. (6) ($n = 20$). Marker color gives the relative error of the production double-precision evaluation at each (S, h) receipt point; ringed markers are cross-gated by a second certified route (41–42 digits, or the $h = \frac{1}{2}$ closed form at ≈ 60), unringed points rest on the oracle’s own ≥ 55 -digit self-checks. Errors reach 2.3% at strong purifying selection with low dominance — a thousand times the tool’s semidominant kernel accuracy.

and the refit is the natural next step.

7 Discussion

What is new here, and what is not. The closed form Eq. (2) is not new; it is folklore, in print at least twice, and in production code twice. The contiguity relations are Gauss’s. What is new: the conditioning analysis that explains why this problem class defeated double precision; the exact $\mathbb{Q}(e^{-S})$ solution of the whole SFS vector and its derivative stack; certified evaluation at every scale of applied interest, including biobank sample sizes; the certified DFE-mixing layer with exact parameter gradients; the audit of the production stack with parameter-level refits; and the entire-deformation structure and first error-controlled evaluation of the dominance spectrum. The negatives are as important: the default deleterious-DFE pipelines of *FITdadi* (current release) and FASTDFE are parameter-level sound at current data sizes — we verified this end to end — and nothing in this paper impugns published deleterious-DFE point estimates that used those defaults within their advertised ranges. The published results that do deserve a recheck are specific: positive-selection quantities (p_b , S_b , α) from POLYDFE model C at $n \gtrsim 90$ haplotypes, and DFE scale and tail-mass statements that passed through *FITdadi*’s 2017 tail-drop rule, particularly for steep DFEs. Both channels come with a certified drop-in replacement here.

Scope and limitations. The certified layer is the equilibrium PRF selection integral and the DFE mixture over it. Real pipelines place nonequilibrium demography upstream (*FITdadi* through the diffusion solver, POLYDFE and FASTDFE through nuisance multipliers); we certify the selection

layer and, where the tools use nuisance-based demography corrections, the full likelihood conditional on them, but we do not certify a demographic model. The dominance treatment is at present for the stationary spectrum only. Our GSL findings are for the version POLYDFE pins; the arithmetic failure is a property of that library version, but the crash-versus-silent dichotomy and the parity dependence are properties of POLYDFE’s own design. Vogl and Bergman (2015) [9], which develops maximum-likelihood machinery for the adjacent boundary-mutation spectrum, contains exactly what our scoping anticipated, now confirmed against the text: an analytic score equation in γ for that spectrum (their Eq. (42), reduced in the large- M limit to a polynomial-root problem, their Appendix A), solved numerically in double precision with no error control; their own discussion states that for the full Wright–Fisher diffusion with directional selection “no analytically tractable results are possible.” Analytic differentiation of a Kummer-family likelihood was therefore known in the adjacent boundary-mutation model, and we credit it; what this paper supplies beyond it is the exact derivative stack to all orders for the PRF selection spectrum and its DFE mixtures, with certified evaluation. The production-tool claims of Section 4 are unaffected.

It would be interesting to extend the certified layer beneath the demography: the time-dependent PRF density solves a forward equation whose moment systems again close on hypergeometric-type towers (finite ladders of special-function recurrences like the ones used here), and a certified treatment there would remove the last uncertified layer of these pipelines. It would also be interesting to take the dominance structure further: the heat relation makes Eq. (6) a two-variable holonomic system of low rank, and transporting in $\tilde{q}$ from the Kummer line is likely the right production evaluator for recessive lethals, where the series pad grows. The tools, gates, and every number in this paper are reproducible from the artifact repository. [\[PENDING: repository URL + artifact SHA list.\]](#)

Acknowledgments

AI-assistance disclosure: Claude (Anthropic) assisted with the numerical work, the literature and code audits, and drafting, under the author’s direction; the author takes full responsibility for the correctness of the work.

References

- [1] S. A. Sawyer and D. L. Hartl, “Population genetics of polymorphism and divergence,” *Genetics* **132**, 1161 (1992).
- [2] P. Tataru, M. Mollion, S. Glémin, and T. Bataillon, “Inference of distribution of fitness effects and proportion of adaptive substitutions from polymorphism data,” *Genetics* **207**, 1103 (2017).
- [3] P. Tataru *et al.*, File S1 of Ref. [2], §2.
- [4] B. Y. Kim, C. D. Huber, and K. E. Lohmueller, “Inference of the distribution of selection coefficients for new nonsynonymous mutations using large samples,” *Genetics* **206**, 345 (2017).
- [5] R. N. Gutenkunst, R. D. Hernandez, S. H. Williamson, and C. D. Bustamante, “Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data,” *PLoS Genetics* **5**, e1000695 (2009).

- [6] J. Sendrowski and T. Bataillon, “fastDFE: fast and flexible inference of the distribution of fitness effects,” *Mol. Biol. Evol.* **41**, msae070 (2024).
- [7] D. Živković, M. Steinrücken, Y. S. Song, and W. Stephan, “Transition densities and sample frequency spectra of diffusion processes with selection and variable population size,” *Genetics* **200**, 601 (2015).
- [8] C. Vogl and F. Clemente, “The allele-frequency spectrum in a decoupled Moran model with mutation, drift, and directional selection, assuming small mutation rates,” *Theor. Popul. Biol.* **81**, 197 (2012).
- [9] C. Vogl and J. Bergman, “Inference of directional selection and mutation parameters assuming equilibrium,” *Theor. Popul. Biol.* **106**, 71 (2015).
- [10] C. D. Bustamante, J. Wakeley, S. Sawyer, and D. L. Hartl, “Directional selection and the site-frequency spectrum,” *Genetics* **159**, 1779 (2001).
- [11] S. Williamson, A. Fledel-Alon, and C. D. Bustamante, “Population genetics of polymorphism and divergence for diploid selection models with arbitrary dominance,” *Genetics* **168**, 463 (2004).
- [12] A. R. Boyko, S. H. Williamson, A. R. Indap, J. D. Degenhardt, R. D. Hernandez, K. E. Lohmueller, M. D. Adams, S. Schmidt, J. J. Sninsky, S. R. Sunyaev, T. J. White, R. Nielsen, A. G. Clark, and C. D. Bustamante, “Assessing the evolutionary impact of amino acid mutations in the human genome,” *PLoS Genetics* **4**, e1000083 (2008).
- [13] P. D. Keightley and A. Eyre-Walker, “Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies,” *Genetics* **177**, 2251 (2007).
- [14] N. Galtier, “Adaptive protein evolution in animals and the effective population size hypothesis,” *PLoS Genetics* **12**, e1005774 (2016).
- [15] H. J. Barton and K. Zeng, “New methods for inferring the distribution of fitness effects for INDELs and SNPs,” *Mol. Biol. Evol.* **35**, 1536 (2018).
- [16] J. Jouganous, W. Long, A. P. Ragsdale, and S. Gravel, “Inferring the joint demographic history of multiple populations: beyond the diffusion approximation,” *Genetics* **206**, 1549 (2017).
- [17] Y. S. Song and M. Steinrücken, “A simple method for finding explicit analytic transition densities of diffusion processes with general diploid selection,” *Genetics* **190**, 1117 (2012).
- [18] C. D. Huber, B. Y. Kim, C. D. Marsden, and K. E. Lohmueller, “Determining the factors driving selective effects of new nonsynonymous mutations,” *Proc. Natl. Acad. Sci. USA* **114**, 4465 (2017).
- [19] C. D. Huber, A. Durvasula, A. M. Hancock, and K. E. Lohmueller, “Gene expression drives the evolution of dominance,” *Nat. Commun.* **9**, 2750 (2018).
- [20] F. W. J. Olver, “Numerical solution of second-order linear difference equations,” *J. Res. Nat. Bur. Standards* **71B**, 111 (1967).

- [21] F. Johansson, “Arb: efficient arbitrary-precision midpoint-radius interval arithmetic,” IEEE Trans. Comput. **66**, 1281 (2017).
- [22] M. Galassi *et al.*, *GNU Scientific Library Reference Manual*, 3rd ed. (2009), ISBN 0954612078; <https://www.gnu.org/software/gsl/>.
- [23] The mpmath development team, *mpmath: a Python library for arbitrary-precision floating-point arithmetic (version 1.3.0)*, 2023; <https://mpmath.org/>.
- [24] *NIST Digital Library of Mathematical Functions*, Release 1.2.7 of 2026-06-15, F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain, eds.; <https://dlmf.nist.gov/>.