BESTish: A Diffusion-Approximation Framework for Inferring Selection and Mutation in Clonal Hematopoiesis

Ren-Yi Wang[1,2,*], Khanh N. Dinh[2,*], Keito Taketomi[2,3], Guodong Pang[4], Katherine Y. King[6,7], and Marek Kimmel[1,5]

[1]Department of Statistics, Rice University, Houston, TX, USA

[2]Irving Institute for Cancer Dynamics and Department of Statistics, Columbia University, New York, NY, USA

[3]Department of Surgery, University of Cambridge, Cambridge, England, UK

[4]Department of Computational Applied Mathematics and Operations Research, Rice University, Houston, TX, USA

[5]Department of Systems Biology and Engineering, Silesian University of Technology, Gliwice, Poland

[6] Division of Infectious Diseases, Department of Pediatrics, Baylor College of Medicine, Houston, TX 77030, USA

[7]Center for Cell and Gene Therapy, Baylor College of Medicine, Houston, TX 77030, USA

[*]These authors contributed equally to this work.

January 27, 2026

Abstract

Clonal hematopoiesis (CH) arises when hematopoietic stem cells (HSCs) gain a fitness advantage from somatic mutations and expand, resulting in an increase in variant allele frequency (VAF) over time. To analyze CH trajectories, we develop a state-dependent stochastic model of wild-type and mutant HSCs, in which an environmental parameter $\alpha \in [0, 1]$ regulates death rates and interpolates between homeostatic (Moran-like, $\alpha = 1$) and growth-facilitating ($\alpha < 1$) regimes. Using functional law of large numbers and central limit theorems, we derive explicit mean-field dynamics and a Gaussian–Markov approximation for VAF fluctuations. We show that the mean VAF trajectory has an explicit logistic form determined by selective advantage, while environmental effects affect only the variance and autocovariance structure. Building on these results, we introduce BESTish (Bayesian estimate for selection incorporating scaling-limit to detect mutant heterogeneity), a novel, efficient and accurate Bayesian inference method that can be applied to analyze both cohort-level and longitudinal VAF datasets. BESTish implements the closed-form finite-dimensional distributions that we derive to estimate mutation fitness, mutation rate, and environmental strength for individual CH drivers. When applied to existing CH datasets, BESTish produces consistent mutation fitness inferences across different studies, and estimates CH driver mutation rates in agreement with independent experimental studies. Furthermore, BESTish reveals patient-specific heterogeneity in the selective behavior of recurrent mutations, and identifies variants whose dynamics are compatible with non-homeostatic, growth-facilitating environments. BESTish provides a unified and mechanistic framework for quantifying CH evolution, with potential applications for other biological systems where clonal expansions can be measured.

1 Introduction

Hematopoiesis is a highly dynamic and tightly regulated process in which hematopoietic stem cells (HSCs) continuously regenerate the diverse pool of blood and immune cells necessary to sustain human life, dividing to maintain nearly 10^{13} cells distributed throughout the body [Cosgrove et al., 2021]. Despite their central role, the total number of HSCs is relatively small, typically on the order of 10^4–10^6 cells in humans [Catlin et al., 2011, Cosgrove et al., 2021, Mitchell et al., 2022], with population-based estimates ranging from approximately 50,000 to 200,000 [Lee-Six et al., 2018]. Moreover, only a fraction of these cells are actively engaged in hematopoiesis at any given time: roughly 30% are thought to be productive [Busch et al., 2015, Cosgrove et al., 2021]. Mitchell et al. [2022] further estimated that 20,000–200,000 cells participate in maintaining blood production, a sum that includes long-, intermediate-, and short-term HSCs as well as multipotent progenitors. With age, HSCs inevitably acquire somatic mutations, some of which provide selective growth advantages. Such mutant clones may expand disproportionately and contribute substantially to hematopoiesis, a phenomenon known as *clonal hematopoiesis* (CH).

Once considered a benign feature of aging, CH is now recognized as a pervasive and clinically relevant phenomenon that bridges somatic evolution, aging, and disease. Large-scale sequencing studies have revealed that CH becomes nearly ubiquitous in older adults and is driven by recurrent mutations in *DNMT3A*, *TET2*, *ASXL1*, splicing factors including *SF3B1*, *SRSF2*, *U2AF1*, and others [Watson et al., 2020, Fabre et al., 2022, Kar et al., 2022]. These mutations confer fitness advantages to HSCs, leading to exponential clonal expansions whose prevalence and growth rates vary with age.

Studying CH provides a quantitative *in vivo* model for understanding how mutation, selection, and drift shape somatic evolution [Watson et al., 2020]. It also illuminates how aging remodels the hematopoietic landscape and predisposes individuals to disease. Faster-growing clones, particularly those carrying *SRSF2* or *U2AF1*

mutations, are associated with increased risk of acute myeloid leukemia and cardiovascular disorders [Fabre et al., 2022]. Environmental and inflammatory factors further modulate CH dynamics, as discussed in [Florez et al., 2022, Hormaechea-Agulla et al., 2021, Bowman et al., 2018, Winter et al., 2024].

The biological system described above can be mathematically represented as a two-compartment model consisting of wild-type (WT) and clonal hematopoiesis (CH) cells. A substantial body of literature addresses mathematical models of multi-compartment cell proliferation, many of which have been applied to the study of cancer initiation and evolution [Durrett, 2015]. In the context of hematopoiesis, spatial constraints within the bone marrow and nonlinear regulatory mechanisms typically limit mathematical analyses to a small number of compartments, requiring the system to be state-dependent. For example, Getto et al. [2013] examined two deterministic models with nonlinear regulation, focusing on global stability. More recently, we extended these models using diffusion approximation techniques to study their stochastic properties [Wang et al., 2025]. Deriving diffusion approximations for the scaled population dynamics greatly facilitates statistical inference, as finite-dimensional distributions can be efficiently computed numerically. An additional advantage of this approach lies in its flexibility: diffusion approximations can be directly mapped to biologically meaningful quantities derived from the underlying population dynamics. This feature is particularly valuable in the study of clonal hematopoiesis, where empirical observations are typically reported as *variant allele frequency* (VAF), approximately half of the ratio between the CH cell count and the total number of hematopoietic stem cells (HSCs).

We develop BESTish (Bayesian estimate for selection incorporating scaling-limit to detect mutant heterogeneity), a quantitative method to infer the mutation and selection rates for individual mutations from their observed VAFs. BESTish is able to infer from both patient-specific longitudinal data, e.g. in [Fabre et al., 2022], and cohort-level observations such as [McKerrell et al., 2015] and [Coombs et al., 2017]. To our knowledge, BESTish represents the first computational effort that utilizes both data types within a consistent mathematical framework to analyze CH mutations, which enhances its applicability for future studies.

BESTish assumes a state-dependent multi-type branching process to model HSC dynamics. The model interpolates between the Moran process (where the population size is conserved; cf. [Wodarz and Komarova, 2014]) and the pure-birth process (where the population size only increases), with the death rate regulated by a parameter $\alpha \in [0, 1]$ that captures the influence of environmental factors on population growth. We further derive the mean-field approximation of the scaled process as the initial population size becomes large. The mean-field dynamics is deterministic with conserved mass when $\alpha = 1$ and grows exponentially when $\alpha < 1$, which can be used to model dynamics under growth-facilitating conditions (e.g., inflammation). We then recover stochasticity by analyzing the fluctuations around the mean-field dynamics, which can be approximated by a Gauss–Markov process for large initial population size, and we derive explicit expressions for its mean and autocovariance functions, which will serve as a key component in BESTish.

The population-level results culminate in the *variant allele frequency* (VAF) [Dentro et al., 2017], defined as one-half times the proportion of CH cells given that CH mutations are invariably heterozygous. VAF serves as a surrogate for clone size, which is difficult to obtain from biological experiment. We derive explicit expressions for the mean-field approximation of the VAF and characterize the corresponding fluctuation dynamics. Notably, we find that the mean-field trajectory of the VAF follows a logistic curve, independent of α, whereas the dependence on α arises solely through the fluctuation structure. This result indicates that environmental regulation influences only the second-order characteristics of the VAF dynamics, such as the autocovariance function. Hence, the impact of α is difficult to detect when sample sizes are small. In contrast, the fitness advantage of CH mutations can be inferred more readily from the VAF, as it determines the slope and trend of the logistic curve. We also note that Fabre et al. [2022] employed logistic functions to fit VAF trajectories and justified this choice via simulations

under the Wright–Fisher model. In the present work, we derive the logistic curve directly from the underlying population dynamics under more general conditions that allow for population growth.

BESTish utilizes these mathematical results to infer key parameters characterizing the VAF dynamics. The algorithm is based on finite-dimensional distributions derived from the diffusion approximation. BESTish is capable of inferring mutation fitness, mutation rate, and strength of environmental factor from both cohort-based and patient-specific longitudinal data.

2 Results

2.1 HSC dynamics $\mathbf{N}^{(r)}$ modulated by environmental factor α

We assume a branching process for the stem cell dynamics in the bone marrow, beginning with r wild-type (WT) cells and no cells carrying clonal hematopoiesis (CH) mutations. The process is characterized with $\mathbf{N}^{(r)}(t) = \left(N_0^{(r)}(t), N_1^{(r)}(t)\right)^\top$, which are respectively the counts of WT and CH cells at time t, starting from initial condition $\mathbf{N}^{(r)}(0) = (r, 0)$. The time until division for cells of type j is exponentially distributed with rate $\lambda_j > 0$, and WT cells mutate and become CH cells with rate v_0. We assume no further mutation for CH cells, hence $v_1 = 0$. The impact of the environment factor on the stem cell population is characterized by $\alpha \in [0, 1]$. The stochastic dynamics can then be described by the following transitions:

WT cell division:	$N_0^{(r)}(t) \to N_0^{(r)}(t) + 1$	at rate $\lambda_0 N_0^{(r)}(t)$,
WT cell mutation:	$\left(N_0^{(r)}(t), N_1^{(r)}(t)\right) \to \left(N_0^{(r)}(t) - 1, N_1^{(r)}(t) + 1\right)$	at rate $v_0 N_0^{(r)}(t)$,
WT cell death:	$N_0^{(r)}(t) \to N_0^{(r)}(t) - 1$	at rate $\dfrac{\alpha \cdot N_0^{(r)}(t)}{\sum_{k=0}^{1} N_k^{(r)}(t)} \sum_{k=0}^{1} \lambda_k N_k^{(r)}(t)$,
CH cell division:	$N_1^{(r)}(t) \to N_1^{(r)}(t) + 1$	at rate $\lambda_1 N_1^{(r)}(t)$, and
CH cell death:	$N_1^{(r)}(t) \to N_1^{(r)}(t) - 1$	at rate $\dfrac{\alpha \cdot N_1^{(r)}(t)}{\sum_{k=0}^{1} N_k^{(r)}(t)} \sum_{k=0}^{1} \lambda_k N_k^{(r)}(t)$.

We observe that $\alpha = 0$ correspond to a pure-birth process with mutation, and $\alpha = 1$ is analogous to the Moran process, which will be elucidated in the following section. We define the fitness of type j individuals to be $w_j := \lambda_j - v_j$ and call the CH mutation neutral if $w_0 = w_1$ and selective if $w_1 > w_0$. Figure 1A provides a schematic representation of our model, and a rigorous construction of the model with Poisson processes is described in Section A.1 of the Appendix.

In Figures 1B-I, we study the simulated dynamics of our HSC model and the resulting VAF trajectories. Figures 1B–E show 100 simulations of the scaled population dynamics $\left(\overline{N}_0^{(r)}(t), \overline{N}_1^{(r)}(t)\right) := \left(N_0^{(r)}(t), N_1^{(r)}(t)\right) \Big/ r$, starting from an initial population size of $r = 20{,}000$ with mutation rate $v_0 = 5 \cdot 10^{-4}$. When $\alpha = 1$, we observe conservation of the total scaled population, i.e. $\overline{N}_0^{(r)}(t) + \overline{N}_1^{(r)}(t) = 1$, consistent with a homeostatic regime for both neutral mutants (Figure 1B) and selective ones (Figure 1C). In contrast, when $\alpha < 1$, the total population expands exponentially, reflecting growth-facilitating environmental conditions (Figures 1D-E).

Figures 1F–I display the corresponding simulated VAF dynamics, defined as $\overline{P}_1^{(r)} := 0.5 \times \overline{N}_1^{(r)} \Big/ \left(\overline{N}_0^{(r)} + \overline{N}_1^{(r)}\right)$. Notably, the mean-field behavior of VAF is invariant with respect to α, while the influence of the environmental factor is subtle and manifests only in the fluctuation (e.g., smaller variance in Figure 1I compared to Figure 1G).

In each figure, solid curves and and shaded regions represent the theoretical mean-field approximations and 95% confidence bands respectively, to be derived in the following sections. We observe that the simulations are

in agreement with our theoretical results across the parameter space. However, in contrast to simulation-based methods such as approximate Bayesian computation (ABC) [Sisson et al., 2018], BESTish's parameter inference utilizes the likelihood functions that we derive mathematically in section 2.4. This both ensures the accuracy in BESTish's parameter estimation and significantly reduces the runtime, as the computational cost for simulations can be extensive, especially for large cell populations.

2.2 Mean-field approximation for the population dynamics

We next derive a mean-field approximation for the population dynamics $\mathbf{N}^{(r)}(t)$. The procedure is analogous to the classical law of large numbers for random variables, treating stochastic processes as random functions. The mean-field approximation captures the central tendency of the dynamics, allowing us to analyze its properties, including growth rate and stability, via differential equations. General discussions on the convergence of stochastic processes can be found at [Billingsley, 1999, Ethier and Kurtz, 2009, Anderson and Kurtz, 2015, Meleard and Bansaye, 2015].

We define the scaled population dynamics to be $\overline{\mathbf{N}}^{(r)}(t) = \mathbf{N}^{(r)}(t)/r$, starting from $\overline{\mathbf{N}}^{(r)}(0) = (1,0)^\top$. Using the functional law of large numbers (FLLN), we show in Proposition 1 in the Appendix that as $r \to \infty$, $\overline{\mathbf{N}}^{(r)}(t)$ converges almost surely to a deterministic function $\overline{\mathbf{N}}(t) = \big(\overline{N}_0(t), \overline{N}_1(t)\big)$, called the *mean-field dynamics* (terminology adapted from [Meleard and Bansaye, 2015]), that follows the initial value problem

$$\begin{aligned}
\frac{d\overline{N}_0(t)}{dt} &= w_0\overline{N}_0(t) - \alpha \cdot \frac{\overline{N}_0(t)}{\sum_{j=0}^{1}\overline{N}_j(t)} \sum_{j=0}^{1}(w_j + v_j)\overline{N}_j(t), \\
\frac{d\overline{N}_1(t)}{dt} &= w_1\overline{N}_1(t) + v_0\overline{N}_0(t) - \alpha \cdot \frac{\overline{N}_1(t)}{\sum_{j=0}^{1}\overline{N}_j(t)} \sum_{j=0}^{1}(w_j + v_j)\overline{N}_j(t),
\end{aligned}$$

starting from $\overline{\mathbf{N}}(0) = (1,0)$, where fitness $w_j := \lambda_j - v_j$ defines the growth rate of cell type j's population. Details can be found in Proposition 1. Observe that

$$\frac{d}{dt}\left(\overline{N}_0(t) + \overline{N}_1(t)\right) = (1-\alpha)[(w_0 + v_0)\overline{N}_0(t) + w_1\overline{N}_1(t)].$$

Hence, $\alpha = 1$ leads to the conservation of total population size, where the model is analogous to the Moran process with mutation [Wodarz and Komarova, 2014].

We consider two parameter regimes, one where the CH mutation is neutral ($w_0 = w_1$) and the mean-field dynamics follows

$$\begin{aligned}
\overline{N}_0(t) &= (1 + v_0 t)^{-\alpha}\, e^{(1-\alpha)w_0 t}, \\
\overline{N}_1(t) &= v_0 t \cdot \overline{N}_0(t).
\end{aligned} \tag{1}$$

The other case corresponds CH cells are selective ($w_1 > w_0$) where the dynamics follows

$$\begin{aligned}
\overline{N}_0(t) &= \left(1 + \frac{v_0}{w_1 - w_0}(e^{(w_1 - w_0)t} - 1)\right)^{-\alpha} e^{(1-\alpha)w_0 t}, \\
\overline{N}_1(t) &= \frac{v_0}{w_1 - w_0}\left(e^{(w_1 - w_0)t} - 1\right)\overline{N}_0(t).
\end{aligned} \tag{2}$$

Note that $\overline{\mathbf{N}}$ in the neutral regime can be obtained from taking the limit $w_1 \to w_0$ in Equation (2).

2.3 Stochastic fluctuation around the mean-field dynamics

In the classical central limit theorem, one takes the scaled difference between the scaled sum of random variables and the mean to recover randomness. Analogously, we center the scaled population dynamics $\overline{\mathbf{N}}^{(r)}$ with respect

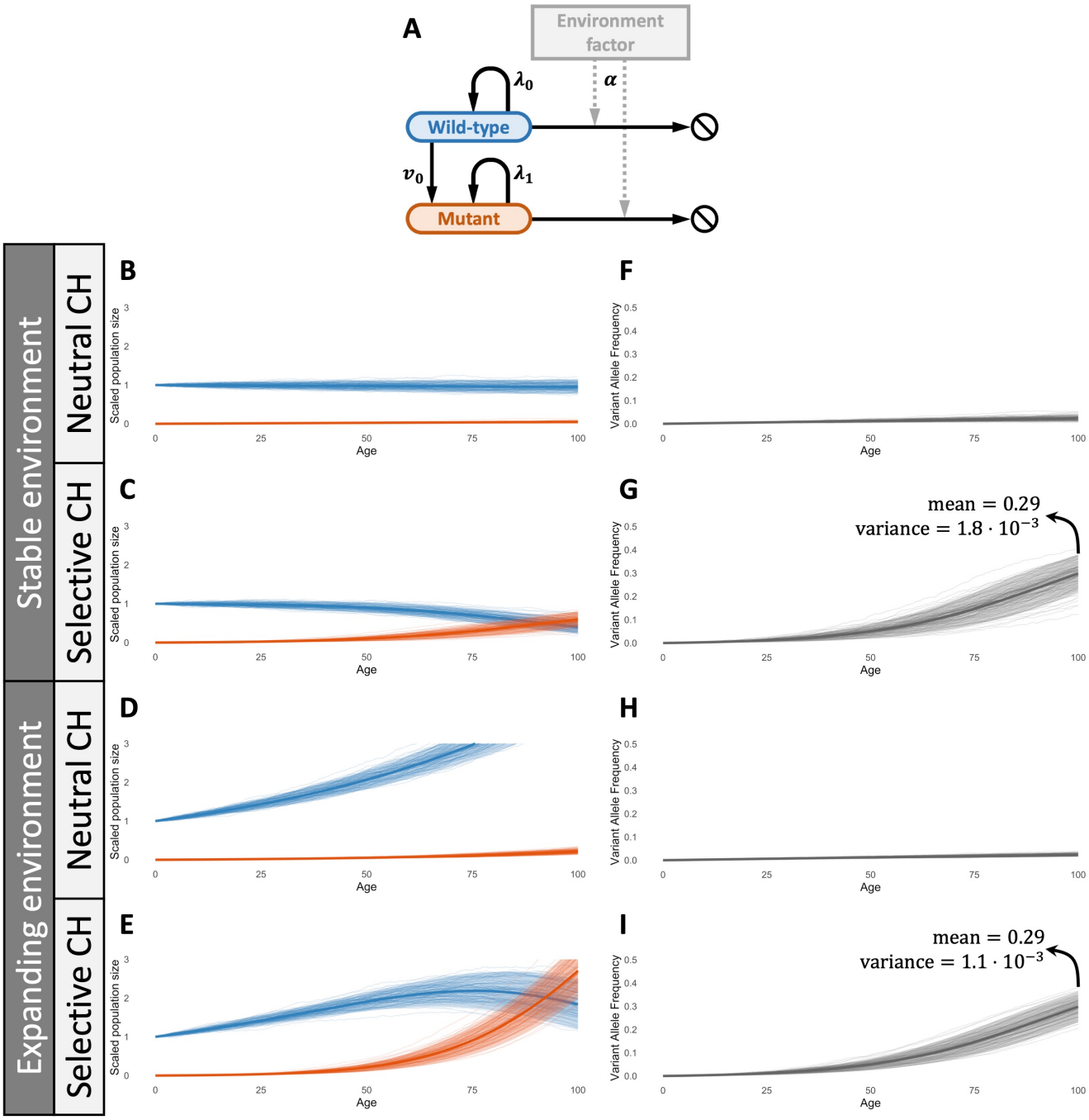


Figure 1: **A**: Schematic for the hematopoietic model with environment factor underlying BESTish. The branching process starts with r wild-type cells, which divide at rate λ_0 and mutate at rate v_0. The mutant cells divide at rate λ_1 and cannot mutate (i.e., $v_1 = 0$). The death rates for both cell types depend on the environment factor, characterized by α. **B-I**: Trajectories of scaled population sizes (**B-E**; blue is wild-type, red is mutant) and corresponding variant allele frequency (VAF; **F-I**) in stable ($\alpha = 1$; **B, C, F, G**) and expanding environments ($\alpha = 0.985$; **D, E, H, I**), where the CH mutant is neutral ($w_0 = w_1 = 1$; **B, F, D, H**) or selective ($w_0 = 1, w_1 = 1.05$; **C, G, E, I**). Other parameters: $v_0 = 5 \cdot 10^{-4}$, $r = 20{,}000$. In each plot, expected values (thick lines; Eqs. (1), (2), (7)) and predicted 95%CI regions (shaded areas; Eqs. (6), (10)) are compared against simulations (thin lines).

to the mean-field dynamics $\overline{\mathbf{N}}$ and scale up the difference. The resulting limiting dynamics $\widehat{\mathbf{N}}$ obtained by the functional central limit theorem (FCLT) is called the *fluctuation dynamics*. To outline the procedure, we define the FCLT-scaled dynamics as

$$\widehat{\mathbf{N}}^{(r)} = \sqrt{r}(\overline{\mathbf{N}}^{(r)} - \overline{\mathbf{N}}); \quad \widehat{\mathbf{N}}^{(r)}(0) = \mathbf{0} \text{ for all } r. \tag{3}$$

We show in Proposition 2 (Appendix) that as $r \to \infty$, $\widehat{\mathbf{N}}^{(r)}$ converges to a Gauss-Markov process $\widehat{\mathbf{N}}$ in distribution. Since $\widehat{\mathbf{N}}$ is a Gauss-Markov process, it can be characterized by its mean function $\mathbf{m}$ and autocovariance function ρ (Karatzas and Shreve [2014]). That is,

$$\widehat{\mathbf{N}} \sim GM(\mathbf{m}, \rho). \tag{4}$$

The variance function $V(t) := \rho(t,t)$ provides insight into the magnitude of fluctuation of population dynamics $\overline{\mathbf{N}}^{(r)}$ when r is large. Specificlly,

$$V_{1,1}(t) = Var(\widehat{N}_0(t)); \quad V_{1,2}(t) = Cov(\widehat{N}_0(t), \widehat{N}_1(t)); \quad V_{2,2}(t) = Var(\widehat{N}_1(t)),$$

and the autocorrelation function $\rho(s,t)_{i,j}/\sqrt{V_{i,i}(s)V_{j,j}(t)}$ encodes self-dependency and range of dependence of $\overline{\mathbf{N}}^{(r)}$. See Wang et al. [2025] for analysis based on convergence of Gaussian processes.

When the initial population size r is large, we can decompose the FLLN-scaled dynamics $\overline{\mathbf{N}}^{(r)}$ as the mean-field dynamics superposed with scaled fluctuation. Since $\widehat{\mathbf{N}}^{(r)} \approx \widehat{\mathbf{N}}$ for large r, by Eq.(3), we have

$$\overline{\mathbf{N}}^{(r)} \approx \overline{\mathbf{N}} + \frac{1}{\sqrt{r}}\widehat{\mathbf{N}} \sim GM(\overline{\mathbf{N}}, \frac{1}{r}\rho). \tag{5}$$

The mean curve and confidence bands for $\overline{\mathbf{N}}^{(r)}$ in Figures 1A,C,D,E are computed by Eq.(5) and its direct consequence Eq.(6) below:

$$\begin{aligned} &\mathbb{P}\left(\overline{N}_0(t) - 1.96\frac{V_{1,1}(t)}{r} \le \overline{N}_0^{(r)}(t) \le \overline{N}_0(t) + 1.96\frac{V_{1,1}(t)}{r}\right) \approx 0.95; \\ &\mathbb{P}\left(\overline{N}_1(t) - 1.96\frac{V_{2,2}(t)}{r} \le \overline{N}_1^{(r)}(t) \le \overline{N}_1(t) + 1.96\frac{V_{2,2}(t)}{r}\right) \approx 0.95. \end{aligned} \tag{6}$$

2.4 Temporal dynamics of the variant allele frequency of CH cells

The variant allele frequency (VAF) dynamics of a given CH mutation is defined by $\overline{P}_1^{(r)} := 0.5 \times \bar{N}_1^{(r)}/(\bar{N}_0^{(r)} + \bar{N}_1^{(r)})$ (cf. Dentro et al. [2017]). Using the continuous mapping theorem, we show that $\overline{P}_1^{(r)}$ converges almost surely to $\overline{P}_1 = 0.5 \times \bar{N}_1/(\bar{N}_0 + \bar{N}_1)$ (Theorem 1, Appendix), which can be simplified to

$$\overline{P}_1(t) = \begin{cases} \frac{1}{2} \cdot \frac{v_0 t}{1+v_0 t}, & \text{if } w_0 = w_1, \\ \frac{1}{2} \cdot \frac{v_0[e^{(w_1-w_0)t}-1]}{w_1-w_0+v_0[e^{(w_1-w_0)t}-1]}, & \text{if } w_1 > w_0. \end{cases} \tag{7}$$

Thus, the mean-field VAF is a logistic curve when the CH mutation confers selective advantage, in accordance to Wright-Fisher simulations in [Fabre et al., 2022]. Our model provides a theoretical derivation for the authors' observations and generalizes to scenarios with expanding population sizes (when $\alpha < 1$).

The steady state for $\overline{P}_1$ is 1/2 for both cases where the CH mutation is neutral ($w_0 = w_1$) or selective ($w_1 > w_0$). The asymptotic rate of convergence to the steady state is $\mathcal{O}(1/t)$ for the neutral scenario. When the CH mutation is selective, $\overline{P}_1$ converges to 1/2 much faster, at the asymptotic rate $\mathcal{O}\left(e^{-(w_1-w_0)t}\right)$. Hence, it is much harder for neutral mutations to be fixed in a large population. Moreover, $\overline{P}_1$ is independent of the environmental parameter α, whose effects manifest only at the level of fluctuations. This indicates that VAF is not an ideal surrogate for clone size when the primary objective is to study clonal expansion.

We then employ the delta method to show that as $r \to \infty$, $\widehat{P}_1^{(r)} = \sqrt{r}\left(\overline{P}_1^{(r)} - \overline{P}_1\right)$ converges in finite-dimensional distributions to $\widehat{P}_1$ (Theorem 2, Appendix), where

$$\widehat{P}_1 \coloneqq \frac{1}{\overline{N}_0 + \overline{N}_1}\begin{bmatrix} -\overline{P}_1 & \overline{P}_0 \end{bmatrix} \cdot \hat{\mathbf{N}} \sim GM(0, K).$$

As a result, the autocovariance function K for $\widehat{P}_1$ is

$$K(s,t) \coloneqq Cov(\widehat{P}_1(s), \widehat{P}_1(t)) = \frac{1}{\overline{N}_0(s) + \overline{N}_1(s)}\begin{bmatrix} -\overline{P}_1(s) & \overline{P}_0(s) \end{bmatrix} \rho(s,t) \begin{bmatrix} -\overline{P}_1(t) \\ \overline{P}_0(t) \end{bmatrix} \frac{1}{\overline{N}_0(t) + \overline{N}_1(t)}.$$

The variance of $\widehat{P}_1$ is then

$$W(t) \coloneqq K(t,t) = \frac{1}{\left(\overline{N}_0(t) + \overline{N}_1(t)\right)^2}\begin{bmatrix} -\overline{P}_1(t) & \overline{P}_0(t) \end{bmatrix} V(t) \begin{bmatrix} -\overline{P}_1(t) \\ \overline{P}_0(t) \end{bmatrix}. \tag{8}$$

We note that the mean-field VAF $\overline{P}_1$ is independent of α, while the fluctuation dynamics $\widehat{P}_1$ depends on α. Similar to Section 2.3, we decompose $\overline{P}_1^{(r)}$ into the deterministic mean-field dynamics and scaled fluctuation for large r:

$$\overline{P}_1^{(r)} \approx \overline{P}_1 + \frac{1}{\sqrt{r}}\widehat{P}_1 \sim GM(\overline{P}_1, \frac{1}{r}K). \tag{9}$$

Illustrative longitudinal mean and confidence bands in Figures 1F-I are computed according to Eq. (9) and its direct consequence:

$$\mathbb{P}\left(\overline{P}_1(t) - 1.96\frac{W(t)}{r} \le \overline{P}_1^{(r)}(t) \le \overline{P}_1(t) + 1.96\frac{W(t)}{r}\right) \approx 0.95. \tag{10}$$

Eq. (9) thus provides us with approximate finite-dimensional distributions for the VAF trajectories, which play a major role in BESTish. To outline the algorithm, we next derive finite-dimensional distributions for both cohort and longitudinal datasets.

2.5 A novel algorithm to infer CH mutation parameters from different experimental settings

BESTish estimates the parameters w_1, v_0, and α from both cohort-based and patient-specific longitudinal datasets. Since all theoretical results in this work rely on the asymptotic regime, we assume sufficiently large initial population size r (typically in the range of 10^4–10^6 cells). Finite-dimensional distributions for inference are derived using Eq. (9).

For a cohort dataset $\{(t_i, y_i)\}_{i=1}^n$ of size n, where the CH mutation is observed with VAF $= y_i$ at age t_i in individual i, the independence across individuals implies that the joint distribution at times $\{t_i\}_{i=1}^n$ follows

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \overline{P}_1(t_1) \\ \vdots \\ \overline{P}_1(t_n) \end{bmatrix}, \frac{1}{r}\begin{bmatrix} K(t_1,t_1) & & \\ & \ddots & \\ & & K(t_n,t_n) \end{bmatrix}\right), \tag{11}$$

where all off-diagonal entries in the covariance matrix are zero due to the inter-individual independence.

Conversely, for a longitudinal dataset $\{(t_i, y_i)\}_{i=1}^n$ where the CH mutation is observed with VAF $= y_i$ at age t_i in the same individual, the temporal dependence must be retained. In this case, Eq. (9) yields

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \overline{P}_1(t_1) \\ \vdots \\ \overline{P}_1(t_n) \end{bmatrix}, \frac{1}{r}\begin{bmatrix} K(t_1,t_1) & \cdots & K(t_1,t_n) \\ \vdots & \ddots & \vdots \\ K(t_n,t_1) & \cdots & K(t_n,t_n) \end{bmatrix}\right). \tag{12}$$

The likelihood functions $\mathcal{L}_{\text{cohort}}(\theta|\{(t_i, y_i)\}_{i=1}^n)$ and $\mathcal{L}_{\text{time-series}}(\theta|\{(t_i, y_i)\}_{i=1}^n)$ are defined from Eqs.(11) and (12) for a parameter set θ using observed VAFs from cohort and time-series datasets, respectively. BESTish infers mutation rate in the logarithmic scale, hence $\theta = (\theta_1, \theta_2, \theta_3, \theta_4) = (w_0, w_1, \log_{10}(v_0), \alpha)$.

We assume that the support of the prior distribution $\pi(\theta)$ is bounded, i.e. $\theta_i \in [\theta_i^{\min}, \theta_i^{\max}]$. BESTish first divides the range for each parameter i into n_i bins, where bin b is $\left[\theta_i^{\min} + b \cdot \delta_i, \theta_i^{\min} + (b+1)\delta_i\right)$. The posterior probability for θ in a given bin (b_1, b_2, b_3, b_4) is then

$$\begin{aligned}\mathbb{P}(\theta \in (b_1, b_2, b_3, b_4)|\{(t_i, y_i)\}_{i=1}^n) &\propto \int_{\theta_4^{\min}+b_4\cdot\delta_4}^{\theta_4^{\min}+(b_4+1)\delta_4} \cdots \int_{\theta_1^{\min}+b_1\cdot\delta_1}^{\theta_1^{\min}+(b_1+1)\delta_1} \pi(\theta)\, \mathcal{L}(\theta|\{(t_i, y_i)\}_{i=1}^n)\, d\theta_1 \cdots d\theta_4 \\ &\approx (\delta_1\delta_2\delta_3\delta_4) \times \pi\left(\theta_1^{\min} + \left(b_1 + \tfrac{1}{2}\right)\delta_1, \ldots, \theta_4^{\min} + \left(b_4 + \tfrac{1}{2}\right)\delta_4\right) \\ &\quad \times \mathcal{L}\left(\left(\theta_1^{\min} + \left(b_1 + \tfrac{1}{2}\right)\delta_1, \ldots, \theta_4^{\min} + \left(b_4 + \tfrac{1}{2}\right)\delta_4\right)\middle|\{(t_i, y_i)\}_{i=1}^n\right)\end{aligned}$$

given that the bin sizes $\delta_1, \ldots, \delta_4$ are small enough, where $\mathcal{L}$ is the appropriate likelihood function for each dataset. BESTish thus computes the posterior probabilities across the $n_1 \times \cdots \times n_4$ bins, and combines them into the joint posterior distribution across the parameter space.

2.6 BESTish reliably infers mutation rate and selective advantage of CH mutations across different data types

We apply BESTish to analyze clonal hematopoiesis data from longitudinal measurements in [Fabre et al., 2022] and cohort-level observations in [McKerrell et al., 2015] and [Coombs et al., 2017]. The model we have employed thus far describes the HSC dynamics within the bone marrow, however the available measurements are derived from peripheral blood samples. Therefore, following Robertson et al. [2022], we assume that the number of differentiated blood cells generated by a given HSC clone is proportional to the size of that clone.

We analyze mutations observed in at least 3 patients in [Fabre et al., 2022] where the variant is measured with VAF > 0 at 3 or more distinct time points. Twenty variants in ten genes with longitudinal data are thus studied: *DNMT3A* ($n = 7$ variants), *TET2* ($n = 3$), *SF3B1* ($n = 2$), *SRSF2* ($n = 2$), *CBL* ($n = 1$), *GNB1* ($n = 1$), *IDH2* ($n = 1$), *JAK2* ($n = 1$), *PPM1D* ($n = 1$) and *U2AF1* ($n = 1$). Additionally, cohort data in [McKerrell et al., 2015] and [Coombs et al., 2017] where a variant is observed in ≥ 8 patients are also incorporated in our analysis. This includes *DNMT3A-R882H* (present in $n = 30$ patients in [McKerrell et al., 2015] and $n = 12$ in [Coombs et al., 2017]), *JAK2-V617F* ($n = 25$ in [McKerrell et al., 2015]), *SF3B1-K700E* ($n = 8$ in [McKerrell et al., 2015]) and *SF3B1-K666N* ($n = 8$ in [McKerrell et al., 2015]).

Catlin et al. [2011], Watson et al. [2020], and Mitchell et al. [2022] estimated that HSCs undergo self-renewal approximately once per year, therefore we fix $w_0 = 1$. For the initial HSC population size, we draw on the estimated number of active HSCs contributing to blood production in [Busch et al., 2015, Cosgrove et al., 2021, Mitchell et al., 2022, Komarova et al., 2024] and set $r = 20{,}000$. Mitchell et al. [2022] estimated using ABC that the driver mutation rate is 2×10^{-3} per HSC per year while Watson et al. [2020] estimated the mutation rates to be much lower, on the scale of 10^{-6}. For cohort-level inference, we assume uniform prior distributions for $\theta = (w_1, \log_{10}(v_0), \alpha)$:

$$\begin{aligned} w_1 &\sim \mathrm{U}(1, 1.3) \\ \log_{10}(v_0) &\sim \mathrm{U}(-6, -3) \\ \alpha &\sim \mathrm{U}(0.5, 1) \end{aligned}$$

However, given the sparse observations for individual variants in [Fabre et al., 2022], we fix $\alpha = 1$ for longitudinal data to avoid overfitting. We apply BESTish with 500 bins for w_1 and $\log_{10}(v_0)$, and 100 bins for α.

The posterior distributions of $(w_1, \log_{10}(v_0))$ for variant *DNMT3A-R882H* from time-series data are presented in Figure 2A. Each patient-specific distribution exhibits a negative correlation between w_1 and $\log_{10}(v_0)$. This reflects the uncertainty in estimating parameters in our model. The same patient-specific VAF dynamic can be explained by a spectrum of parameters in BESTish: low mutation rate $\log_{10}(v_0)$ accompanied by high selective advantage w_1, or vice versa.

Comparing BESTish's results reveals that *DNMT3A-R882H* behaves more similarly to a neutral mutation (i.e., low w_1) in five patients, and is more characteristic of a driver event (i.e., high w_1) in patient PD41103. Importantly, VAF simulations assuming the maximum a posteriori (MAP) estimates from each distribution closely match the observed dynamics in [Fabre et al., 2022] (Figure 2B), underlying the accuracy of the inferred patient-specific parameters.

The applications of BESTish for *DNMT3A-R882H* using cohort-level measurements separately from [Coombs et al., 2017] and [McKerrell et al., 2015] also result in VAF dynamics that closely match observed data (Figures 2C-D). The expected values and 95%CI regions, derived in previous sections, assuming the MAP estimates for each cohort are also in strong agreement with simulations. Therefore, the implementation of these mathematical results enables BESTish to uncover the parameter distributions more efficiently compared to simulation-based approaches, which would require creating large-scale numerical experiments to approximate the distributions at much higher time costs.

Remarkably, the independent inferences lead to highly consistent results for $(w_1, \log_{10}(v_0))$ between the two cohorts (Figures 2A), despite differences in both patient ages ([Coombs et al., 2017]: median = 69 years, range $[43, 78]$; [McKerrell et al., 2015]: median = 62 years, range $[25, 82]$) and measured VAFs ([Coombs et al., 2017]: median = 0.072, range $[0.035, 0.265]$; [McKerrell et al., 2015]: median = 0.027, range $[0.008, 0.320]$). This confirms that BESTish is robust to data heterogeneity and can reliably infer parameters characterizing individual CH mutations.

We note that the inferred $(w_1, \log_{10}(v_0))$ from the cohorts reside within the range of patient-specific inferences from [Fabre et al., 2022] (Figures 2A). This supports our view that cohort-based estimates from BESTish represent population-level averages of the same CH variants' diverse behaviors in distinct individuals. Finally, the inferred environment factor α is approximately 1 in both cohorts (Figures 2E), indicating that the cell dynamics associated with *DNMT3A-R882H* are likely homeostatic.

BESTish's results from fitting time-series data for 19 other mutations in our study are shown in Supplementary Figures 1-19. Similar to *DNMT3A-R882H*, the simulated VAF trajectories with MAP estimates from distinct inferences closely match the corresponding patient-specific dynamics. This confirms that BESTish can uncover the shared characteristics of a given mutation, despite little overlap in age and VAF ranges from different samples in many instances.

2.7 Behavior of CH variants as driver or neutral events is patient-specific

BESTish's ability to infer patient-specific parameters characterizing a mutation enables a more detailed analysis of CH variants' diverse behaviors. Figure 3A displays the inferred selection rates for 20 variants across 100 patients included in our study of longitudinal data in [Fabre et al., 2022]. Similar to *DNMT3A-R882H* (Figure 2A), when data in [Coombs et al., 2017] and [McKerrell et al., 2015] is available, the cohort-based inferences consistently fall within the range of estimates derived for individuals (e.g., *JAK2-V617F*, *SF3B1-K666N* and *SF3B1-K700E*,

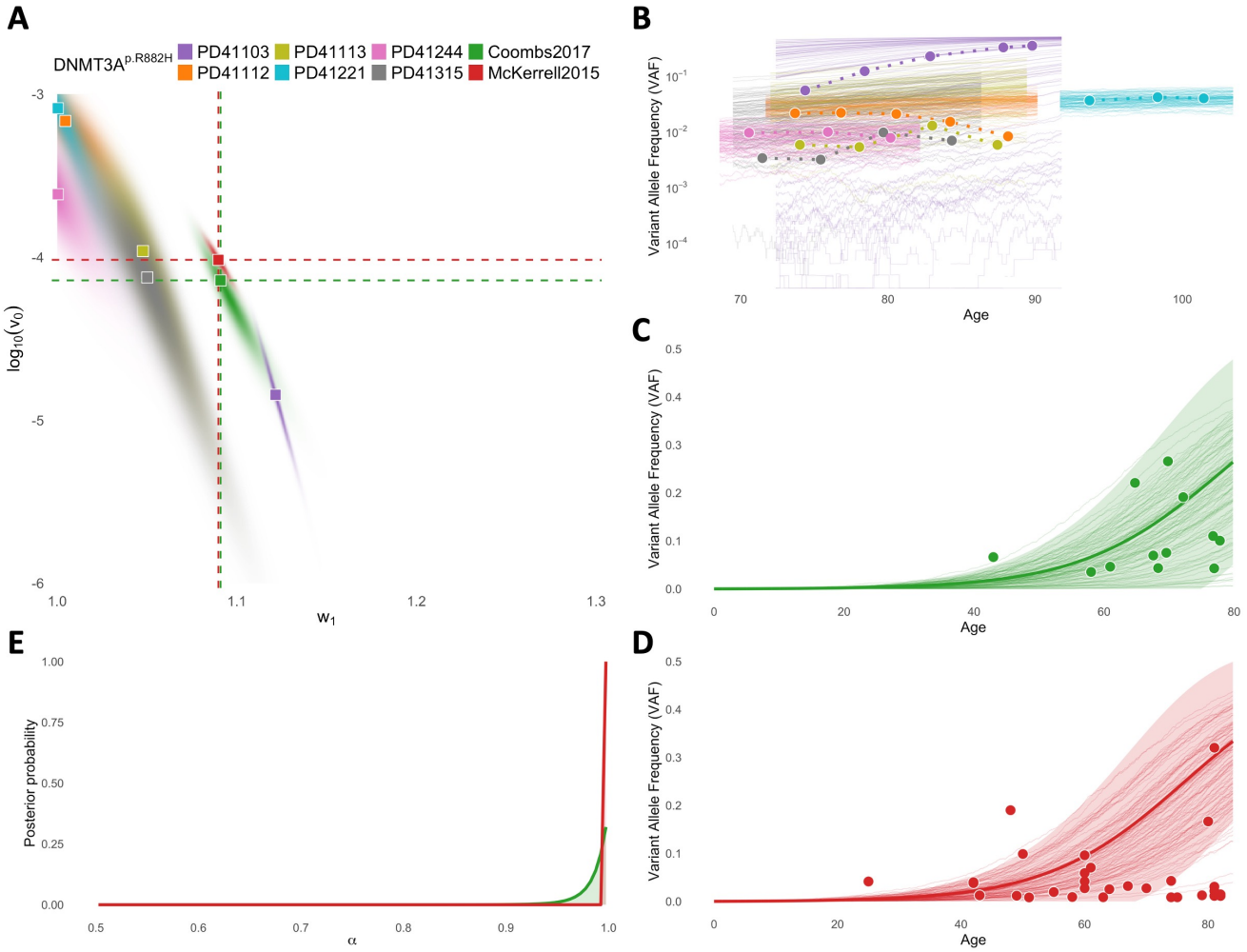


Figure 2: Inference of *DNMT3A-R882H*'s mutation rate and selection rate from cohort samples in [Coombs et al., 2017] ("Coombs2017") and [McKerrell et al., 2015] ("McKerrell2015"), and time-series trajectories in [Fabre et al., 2022] (patient IDs: PD41103, PD411112, PD411113, PD41221, PD41244, PD41315). **A**: Joint posterior distributions for logarithmic mutation rate ($\log_{10}(v_0)$) and selection rate (w_1) for each dataset. Squares indicate maximum a posteriori (MAP) estimates from each distribution. Prior distributions: $\log_{10}(v_0) \sim \mathrm{U}(-6,-3), w_1 \sim \mathrm{U}(1,1.3)$; for "Coombs2017" and "McKerrell2015": $\alpha \sim \mathrm{U}(0.5,1)$. Fixed parameters: $w_0 = 1, r = 2 \times 10^4$, for time-series data: $\alpha = 1$. **B**: 100 simulated variant allele frequency (VAF) trajectories in logscale assuming MAP estimates (thin lines), against VAFs from corresponding patient-specific longitudinal data (circles, connected by dashed lines). For legibility, simulations are truncated to the age range of the corresponding dataset. **C, D**: Expected values (thick lines) and predicted 95%CI regions (shaded areas) for VAF trajectories assuming MAP estimates, against observed VAFs (circles) and simulations with MAP parameters (thin lines), based on data in [Coombs et al., 2017] (**C**) and [McKerrell et al., 2015] (**D**). **E**: Posterior distributions for environment factor α from [Coombs et al., 2017] and [McKerrell et al., 2015]. Colors in **B-E** correspond to datasets in **A**.

Figure 3A). Hence, the inter-patient heterogeneity in BESTish's inferred fitnesses explains the population-level observations for single CH variants.

By delivering patient-specific parameter estimates, BESTish enables sensitive detection of potential CH driver mutations. The results for *DNMT3A-R882H* (Figure 2) suggest that a threshold of $w_1 = 1.1$ provides a reasonable criterion for distinguishing neutral from selective CH variants in our model. For $w_1 < 1.1$, the VAF either remains stable or displays negligible growth over time, consistent with neutral evolution (e.g., PD41112, PD411113, PD41221, PD41244, PD41315; Figure 2B). On the other hand, a high selection rate ($w_1 > 1.1$) corresponds to a substantial VAF increase, indicating an ongoing selective sweep (e.g., PD41103; Figure 2B). We identify 8 variants that are highly selective (i.e., $w_1 > 1.1$) in one or more patients: *DNMT3A-R882H*, *JAK2-V617F*, *SF3B1-K666N*, *SRSF2-P95H*, *SRSF2-P95L*, *TET2-I1873T*, *TET2-I274fs* and *U2AF1-Q84P* (Figure 3A). Among these, two variants were overlooked in the original study [Fabre et al., 2022], likely because the authors estimated mutation fitness across multiple patients simultaneously, an approach that is comparable to BESTish's cohort-based inferences and can only capture the average behaviors of CH variants. The first variant is *I1873T*, determined as likely oncogenic in OncoKB [Chakravarty et al., 2017] and affecting *TET2*, a tumor suppressor and DNA demethylase frequently mutated in hematologic malignancies. The second variant, *P95L* in RNA splicing factor *SRSF2*, is also likely oncogenic [Chakravarty et al., 2017] and has been observed to be a statistically significant hotspot in chronic myeloid leukemia [Zhang et al., 2019].

We then examine the mutation rates inferred from BESTish for the 20 variants in this study (Figure 3B). Despite significant differences in inferred $\log_{10}(v_0)$ among individuals in the time-series dataset for certain mutations, the median mutation rates across longitudinal patients are always within the 95%CI of the cohort-based inferences when data is available (e.g., *DNMT3A-R882H*, *JAK2-V617F*, *SF3B1-K666N* and *SF3B1-K700E*, Figure 3B). Furthermore, the sum of median longitudinal $\log_{10}(v_0)$ across the 20 variants yields the total mutation rate of 1.2×10^{-3}. This figure is remarkably close to the rate of 2×10^{-3} driver mutations per HSC per year inferred by Mitchell et al. [2022], who implemented approximate Bayesian computation to estimate mutation rate from single-cell sequencing data for 10 patients between 0 and 81 years of age. The agreement of the inferred driver mutation rates, despite differences in modeling approaches and experimental contexts, further validates the results from BESTish.

We next assess whether the apparent selective behavior of certain CH variants in individual patients may be attributable to other co-occurring driver mutations. To address this, we tabulate the co-occurrence and mutual exclusion in [Fabre et al., 2022] of the 10 genes in this study in addition to *TP53* and *ASXL1*, two genes that have been observed to be frequently mutated in clonal hematopoiesis [Kar et al., 2022](Figure 3C). The only significant pattern is a mutual exclusion between *DNMT3A* and *ASXL1*. There is no significant co-occurrence between other genes, consistent with previous findings by Kar et al. [2022]. This suggests that the CH variants are likely responsible for the selective sweeps observed in patients where BESTish determines their fitness to be high.

We next investigate how clinical covariates contribute to the mutant-specific intra-patient heterogeneity of fitness observed in Figure 3A. We use two approaches, permutation importance and SHAP, to rank the influence of clinical measurements on patient- and mutant-specific selection rates. Because the data are longitudinal with distinct time points for each patient, we summarize each covariate by its mean level and temporal change rate, the latter defined as the slope of the linear regression fitted to measured values against age (details in Methods).

Notably, the three most important covariates identified by both methods are the change rates of alanine aminotransferase (ALT), C-reactive protein (CRP), and ferritin (Figures 3D-E). ALT and CRP change rates correlate with gene-adjusted patient- and mutant-specific selection rates with statistical significance (p-value

< 0.01, Figures 3F-G), and both proteins are related to liver function and inflammation Wong et al. [2023]. This is consistent with previous findings that CH is associated with elevated pro-inflammatory cytokines, which may act as a driving force for clonal expansion Avagyan and Zon [2023].

We further observe that the mean levels of ALT and CRP are positively correlated with fitness, but these correlations are weaker than those obtained using their change rates. Interestingly, although the three change rates have stronger predictive power for CH mutant fitness, their Spearman rank correlations are negative (Figures 3F-H). This seemingly paradoxical observation requires further investigation, especially given the limited number of individuals with sufficiently dense longitudinal measurements.

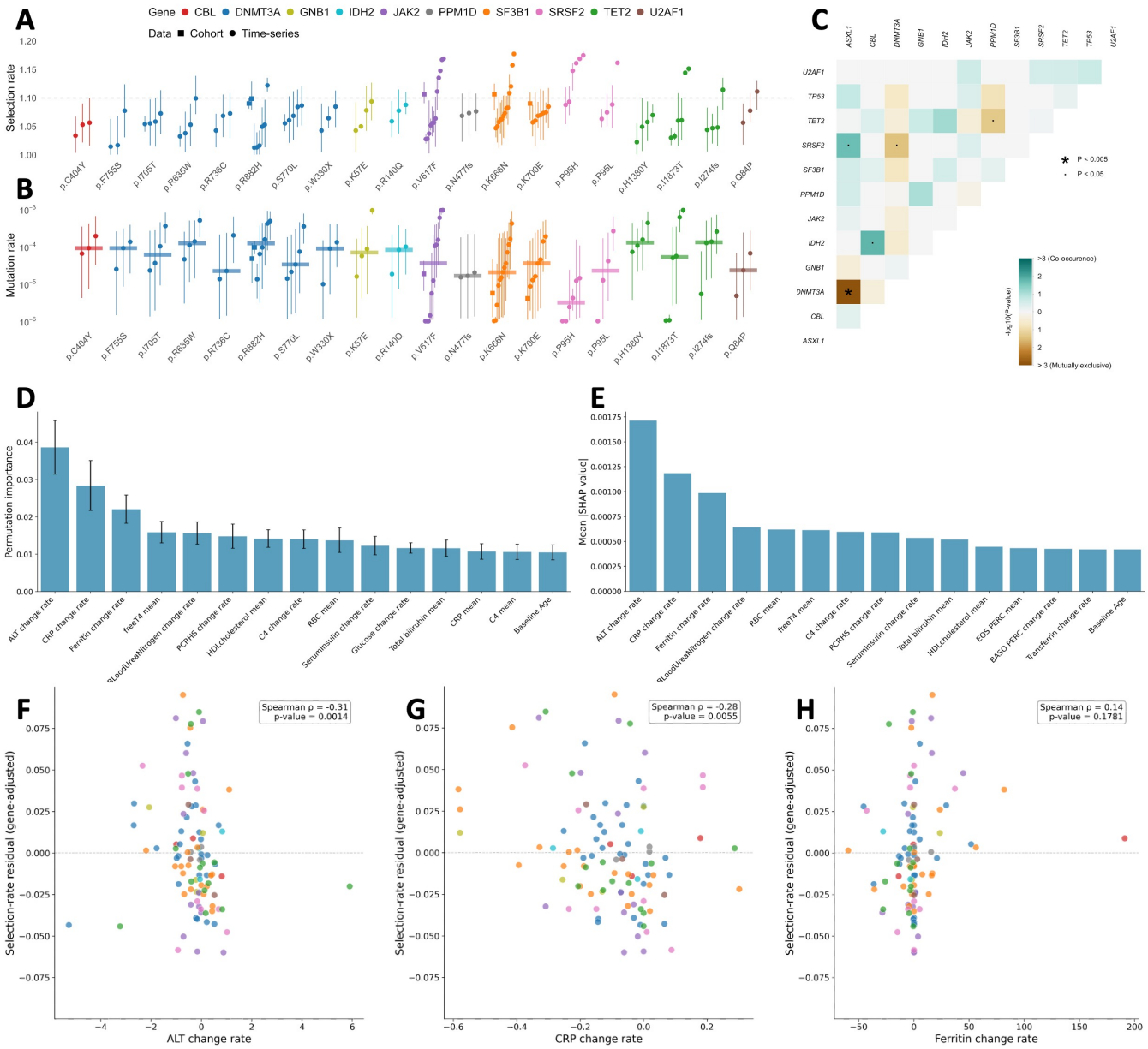


Figure 3: Analysis of mutant-specific intra-patient heterogeneity of fitness advantage inferred from BESTish. **A, B**: Selection rates (**A**) and mutation rates (**B**) inferred from BESTish for 20 frequent CH variants from cohort samples in [McKerrell et al., 2015] and [Coombs et al., 2017] ("Cohort") and longitudinal trajectories in [Fabre et al., 2022] ("Time-series"). Each point represents the median from the posterior distribution of w_1 (**A**) or $\log_{10}(v_0)$ (**B**) inferred from a specific cohort or longitudinal patient, and vertical bars display the 95%CI. Horizontal bars in **B** represent the median mutation rates across longitudinal inferences for each CH variant. **C**: Pairwise association matrix between mutated CH genes, using data from [Fabre et al., 2022]. Color indicates the co-occurrence or mutual exclusion of mutations affecting each pair of genes. Gene list consists of *TP53*, *ASXL1* and those included in **A**. P-values are calculated from Fisher's exact test, with false discovery rate correction applied to account for multiple testing. **D, E**: Contributions of clinical observations in explaining patient- and variant-specific selection rates in **A**, using feature permutation with random forest (**D**) and SHAP (SHapley Additive exPlanations) (**E**). **F, G, H**: Relationship between BESTish's inferred variant- and patient-specific selection rates and change rates of ALT (**F**), CRP (**G**) and ferritin (**H**). The color of each mutation (circle) corresponds to genes in **A**. P-value and ρ from Spearman's rank correlation.

2.8 The impact of the environment factor

Since $\alpha < 1$ implies net growth of the HSC population, an estimated value of α below one suggests that the corresponding mutation may be associated with growth-facilitating conditions. However, when the data are observed only through VAF measurements, α is identifiable primarily at the level of fluctuations, which requires substantially more samples for accurate estimation. This limitation arises because VAF is a one-dimensional summary of an underlying two-dimensional process (wild-type and CH cell counts), and thus inevitably entails a loss of information. For instance, illustrative examples in Figures 1C, G, E, I are simulated with identical parameters except for α. While the cell population progressions exhibit markedly different behaviors (Figures 1C, E), the corresponding VAF trajectories have identical means throughout time and only differ at the level of fluctuations (Figures 1G, I). This observation further highlights that VAF is not an ideal surrogate for true clone sizes.

In Supplementary Figure 3, the posterior distribution of α does not concentrate at 1, suggesting that *SF3B1-K700E* (a splicing-factor mutation) may be associated with growth-facilitating conditions, such as inflammation [Choudhary et al., 2022]. Supplementary Figure 2 shows an even larger deviation of α from 1. When combined with the low estimated mutation rate, this pattern suggests that *SF3B1-K666N* may be linked to hematologic malignancy, consistent with clinical observations showing that K666 mutations tend to arrive late and are strongly associated with high-risk MDS and AML [Bick et al., 2020, Chen et al., 2021]. However, these conclusions remain speculative due to limited sample sizes of the cohort-level datasets. We suggest that future studies based on larger population cohorts may utilize BESTish to infer conclusively the role of environmental impacts associated with specific CH variants.

3 Discussion

BESTish offers three advantages with respect to parameter inference that set it apart from previous efforts [Zhang and Bozic, 2024, Watson et al., 2020, Fabre et al., 2022]. First, it incorporates our mathematical results for the mean and variance of the VAF dynamics. This enables BESTish to derive the distributions for parameters characterizing each CH variant more accurately and efficiently, compared to simulation-based methods. Second, BESTish can analyze data from either cohort-level studies or time-series experiments. This both enhances its applicability for future studies and allows for consistent analyses of the same CH mutations across different studies and experiments. Finally, BESTish is able to estimate parameters from single patients in longitudinal data, facilitating more in-depth evaluations of covariates influencing CH variants' heterogeneous behavior as driver or passenger events in different individuals.

Owing to the flexibility of our framework, the methods developed in this paper can be extended to other biological settings involving mutations and interactions with the microenvironment, such as tumorigenesis. A rich literature has proposed stochastic models of tumor evolution with varying levels of mutational complexity [Durrett and Moseley, 2010, Durrett et al., 2011, Antal and Krapivsky, 2011, Nicholson and Antal, 2019, Johnson et al., 2023, Zhang and Bozic, 2024]. However, most existing models do not explicitly incorporate environmental regulation, and their exact branching-process formulations typically render the derivation of finite-dimensional distributions for temporal dynamics analytically intractable. For example, Zhang et al. [2024] emphasize that in multistage tumorigenesis the microenvironment evolves from being tumor-suppressive to malignancy-supporting. In our framework, such a transition can be naturally captured by allowing the environmental parameter α to depend on time or on the system state. Moreover, the associated diffusion approximation provides tractable

approximate finite-dimensional distributions for the temporal dynamics, thereby substantially simplifying the statistical inference.

Methods

In this section, we outline step-by-step derivations of quantities that will be used in BESTish. Details and proofs will be deferred to the Appendix.

State-dependent branching process model

Let $N_0^{(r)}(t)$ and $N_1^{(r)}(t)$ denote the number of WT (type 0) and CH (type 1) cells at time t with initial population consisting of r WT cells. The rate of division of a type j cell is $\lambda_j > 0$ and the mutation rate is $v_j \geq 0$. Since type 1 individuals cannot further mutate, $v_1 = 0$. All individuals suffer from the same death rate modulated by $\alpha \in [0, 1]$:

$$\frac{\alpha \sum_{j=0}^{1} \lambda_j N_j^{(r)}}{\sum_{j=0}^{1} N_j^{(r)}}.$$

For our model, we define fitness of a type j individual by $w_j \coloneqq \lambda_j - v_j$. We are interested in two parameter regimes:

$$\begin{cases} w_1 = w_0 \text{ (neutral mutation)} \\ w_1 > w_0 \text{ (selective mutation).} \end{cases}$$

Functional law of large numbers

Define the density dynamics by

$$\overline{\mathbf{N}}^{(r)}(t) \coloneqq \frac{\mathbf{N}^{(r)}(t)}{r}, \qquad \overline{\mathbf{N}}^{(r)}(0) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$

Since r represents the initial number of HSCs, it may also be viewed as a proxy for the physical space in which proliferation occurs. Consequently, $\overline{\mathbf{N}}^{(r)}$ can be interpreted as a density, and increasing r corresponds to expanding the effective bone-marrow space.

As $r \to \infty$, Proposition 1 shows that $\overline{\mathbf{N}}^{(r)}$ converges in probability to the deterministic limit $\overline{\mathbf{N}}$, whose components satisfy

$$\begin{aligned} \overline{N}_0'(t) &= w_0\, \overline{N}_0(t) - \alpha \frac{\overline{N}_0(t)}{\overline{N}_0(t) + \overline{N}_1(t)} \sum_{j=0}^{1} (w_j + v_j)\, \overline{N}_j(t), \\ \overline{N}_1'(t) &= w_1\, \overline{N}_1(t) + v_0\, \overline{N}_0(t) - \alpha \frac{\overline{N}_1(t)}{\overline{N}_0(t) + \overline{N}_1(t)} \sum_{j=0}^{1} (w_j + v_j)\, \overline{N}_j(t), \\ &\text{with initial condition } \overline{\mathbf{N}}(0) = (1, 0)^\top. \end{aligned}$$

This is an autonomous system of ODEs, which can be succinctly expressed as $\overline{\mathbf{N}}' = \mathbf{F}(\overline{\mathbf{N}})$.

To solve this system, consider the ratio

$$R_1(t) \coloneqq \frac{\overline{N}_1(t)}{\overline{N}_0(t)}.$$

A direct computation shows that the dynamics of R_1 are independent of α:

$$R_1'(t) = (w_1 - w_0)R_1(t) + v_0.$$

Lemma 1 then gives the explicit solution

$$R_1(t) = \begin{cases} v_0 t, & w_0 = w_1, \\ \dfrac{v_0}{w_1 - w_0}\left(e^{(w_1-w_0)t} - 1\right), & w_0 \neq w_1. \end{cases}$$

Using $\overline{N}_1(t) = R_1(t)\,\overline{N}_0(t)$, we obtain

$$\frac{d}{dt}\ln \overline{N}_0(t) = w_0 - \alpha\frac{(w_0+v_0) + w_1 R_1(t)}{1 + R_1(t)}, \qquad \ln \overline{N}_0(0) = 0.$$

Integrating yields

$$\begin{aligned} \overline{N}_0(t) &= \begin{cases} \left(1 + \frac{v_0}{w_1-w_0}\left(e^{(w_1-w_0)t} - 1\right)\right)^{-\alpha} e^{(1-\alpha)w_0 t}, & \text{if } w_1 > w_0 \\ (1+v_0 t)^{-\alpha}\, e^{(1-\alpha)w_0 t}, & \text{if } w_1 = w_0 \end{cases} \\ \overline{N}_1(t) &= R_1(t)\,\overline{N}_0(t) = \begin{cases} \frac{v_0}{w_1-w_0}\left(e^{(w_1-w_0)t} - 1\right)\overline{N}_0(t) & \text{if } w_1 > w_0 \\ v_0 t \overline{N}_0(t), & \text{if } w_1 = w_0. \end{cases} \end{aligned} \tag{13}$$

Functional central limit theorem

Define the fluctuation dynamics by

$$\widehat{\mathbf{N}}^{(r)} := \sqrt{r}(\overline{\mathbf{N}}^{(r)} - \overline{\mathbf{N}}); \qquad \widehat{\mathbf{N}}^{(r)}(0) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

$\widehat{\mathbf{N}}^{(r)}$ characterizes stochastic fluctuation around the density dynamics. As $r \to \infty$, we show in Proposition 2 that $\widehat{\mathbf{N}}^{(r)}$ converges in distribution to $\widehat{\mathbf{N}}$, which is characterized by a linear diffusion with time-dependent coefficients

$$d\widehat{\mathbf{N}}(t) = \nabla\mathbf{F}(\overline{\mathbf{N}}(t))\widehat{\mathbf{N}}(t)dt + \sigma(t)d\mathbf{B}(t); \qquad \widehat{\mathbf{N}}(0) = 0.$$

Here, $\mathbf{B}$ is a 5−dimensional standard Brownian motion and

$$\sigma^\top(t) = \begin{bmatrix} \sqrt{(w_0+v_0)\overline{N}_0} & 0 \\ -\sqrt{v_0\overline{N}_0} & \sqrt{v_0\overline{N}_0} \\ -\sqrt{\frac{\alpha\overline{N}_0}{\sum_{j=0}^1 \overline{N}_j}\sum_{j=0}^1 (w_j+v_j)\overline{N}_j} & 0 \\ 0 & \sqrt{(w_1+v_1)\overline{N}_1} \\ 0 & -\sqrt{\frac{\alpha\overline{N}_1}{\sum_{j=0}^1 \overline{N}_j}\sum_{j=0}^1 (w_j+v_j)\overline{N}_j} \end{bmatrix}.$$

The first row of σ corresponds to birth, mutate, and death rates for WT cells and the second row corresponds to birth and death rates for CH cells.

Define a 2×2 matrix-valued function Φ by

$$\Phi'(t) = \nabla\mathbf{F}(\overline{N}(t)); \quad \Phi(0) = I_{2\times 2}.$$

Let $\Psi(t) \coloneqq \Phi^{-1}(t)$. In Proposition 2, we also show that $\widehat{\mathbf{N}}$ follows a Gauss-Markov process with mean function $\mathbf{m} \equiv 0$ and autocovariance function

$$\rho(s,t) = \Phi(s)[\int_0^{\min(s,t)} \Psi(u)\sigma(u)\sigma^\top(u)\Psi(u)^\top du]\Phi^\top(t). \tag{14}$$

Define the variance function by $V(t) = \rho(t,t)$, then

$$V'(t) = \nabla\mathbf{F}(\overline{\mathbf{N}}(t))V(t) + [\nabla\mathbf{F}(\overline{\mathbf{N}}(t))V(t)]^\top + \sigma(t)\sigma^\top(t); \qquad V(0) = 0_{2\times 2}. \tag{15}$$

For $s \le t$, we have

$$\rho(s,t) = V(s)\Psi^\top(s)\Phi^\top(t). \tag{16}$$

This allows us to compute the joint distribution for patient-specific longitudinal data efficiently in BESTish.

Variant allele frequency

Define a mapping $\mathbf{g} : \mathcal{D}^2 \to \mathcal{D}$, where $\mathcal{D}$ is the space of functions that are right-continuous with left limits, such that

$$\mathbf{g}(x,y) = \frac{1}{2}\frac{y}{x+y}.$$

The VAF is then defined by

$$\overline{P}_1^{(r)} \coloneqq \mathbf{g}(\overline{\mathbf{N}}^{(r)}).$$

In Theorem 1, we show as $r \to \infty$,

$$\overline{P}_1^{(r)} = \mathbf{g}(\overline{\mathbf{N}}^{(r)}) \to \mathbf{g}(\overline{\mathbf{N}}) \coloneqq \overline{P}_1 \text{ almost surely.}$$

Hence,

$$\overline{P}_1(t) = \begin{cases} \frac{1}{2}\cdot\frac{v_0 t}{1+v_0 t}, & \text{if } w_0 = w_1, \\ \frac{1}{2}\cdot\frac{v_0[e^{(w_1-w_0)t}-1]}{w_1-w_0+v_0[e^{(w_1-w_0)t}-1]}, & \text{if } w_1 > w_0. \end{cases} \tag{17}$$

Define $\widehat{P}_1^{(r)}$ and $\widehat{P}_1$ such that for all $t \ge 0$,

$$\widehat{P}_1^{(r)}(t) \coloneqq \sqrt{r}(\widehat{P}_1^{(r)}(t) - \widehat{P}_1(t)); \quad \widehat{P}_1(t) \coloneqq \nabla\mathbf{h}(\overline{\mathbf{N}}(t))\widehat{\mathbf{N}}(t),$$

where $\mathbf{h} : \mathbb{R}^2 \to \mathbb{R}$ such that

$$\mathbf{h}(x,y) = \frac{1}{2}\frac{y}{x+y}.$$

We show using delta method in Theorem 2 that as $r \to \infty$, the following convergence in finite-dimensional distribution holds:

$$\widehat{P}_1^{(r)} \overset{FDD}{\to} \widehat{P}_1 \sim GM(\mathbf{0}, K). \tag{18}$$

The autocovariance function K for $\widehat{P}_1$ is

$$\begin{aligned} K(s,t) &\coloneqq \nabla\mathbf{h}(\overline{\mathbf{N}}(s))\rho(s,t)\nabla\mathbf{h}^\top(\overline{\mathbf{N}}(t)) \\ &= \frac{1}{\overline{N}_0(s)+\overline{N}_1(s)}\begin{bmatrix} -\overline{P}_1(s) & \overline{P}_0(s) \end{bmatrix}\rho(s,t)\begin{bmatrix} -\overline{P}_1(t) \\ \overline{P}_0(t) \end{bmatrix}\frac{1}{\overline{N}_0(t)+\overline{N}_1(t)}. \end{aligned} \tag{19}$$

Finite-dimensional distributions in BESTish

The key ingredient of our algorithm is the Gaussian–Markov approximation

$$\overline{P}_1^{(r)} \approx GM\left(\overline{P}_1,\ \frac{1}{r}K(\cdot,\cdot)\right), \tag{20}$$

derived from Eq. (18). In other words, $\overline{P}_1^{(r)}$ is (approximately) a Gauss-Markov process with mean function $\overline{P}_1$ and autocovariance kernel K/r. Since the explicit expressions for $\overline{\mathbf{N}}$ and $\overline{P}_1$ are available (Eqs. (13) and (17)), $K(s,t)$ in Eq. (19) can be computed by numerical integrating Eq. (14), which can be computed more efficiently by using Eqs. (15) and (16).

For a cohort dataset $\{(t_i, y_i)\}_{i=1}^n$ of size n, where the CH mutation is observed with VAF $= y_i$ at age t_i in individual i, the independence across individuals implies that the joint distribution at times $\{t_i\}_{i=1}^n$ follows

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \overline{P}_1(t_1) \\ \vdots \\ \overline{P}_1(t_n) \end{bmatrix}, \frac{1}{r}\begin{bmatrix} K(t_1,t_1) & & \\ & \ddots & \\ & & K(t_n,t_n) \end{bmatrix}\right), \tag{21}$$

where all off-diagonal entries in the covariance matrix are zero due to the inter-individual independence.

Conversely, for a longitudinal dataset $\{(t_i, y_i)\}_{i=1}^n$ where the CH mutation is observed with VAF $= y_i$ at age t_i in the same individual, the temporal dependence must be retained. In this case, Eq. (20) yields

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \overline{P}_1(t_1) \\ \vdots \\ \overline{P}_1(t_n) \end{bmatrix}, \frac{1}{r}\begin{bmatrix} K(t_1,t_1) & \cdots & K(t_1,t_n) \\ \vdots & \ddots & \vdots \\ K(t_n,t_1) & \cdots & K(t_n,t_n) \end{bmatrix}\right). \tag{22}$$

Here, the full covariance matrix captures the temporal correlations in the VAF trajectory.

BESTish infers $\theta = (\theta_1, \theta_2, \theta_3, \theta_4) = (w_0, w_1, \log_{10}(v_0), \alpha)$ using the likelihood functions $\mathcal{L}_{\text{cohort}}\left(\theta|\{(t_i, y_i)\}_{i=1}^n\right)$ and $\mathcal{L}_{\text{time-series}}\left(\theta|\{(t_i, y_i)\}_{i=1}^n\right)$, defined based on Eqs. (21) and (22) for population-level and longitudinal data, respectively. Assuming that θ_i is uniformly distributed on $[\theta_i^{\min}, \theta_i^{\max}]$, BESTish discretizes the support for each parameter i into n_i bins, where bin b is $\left[\theta_i^{\min} + b \cdot \delta_i, \theta_i^{\min} + (b+1)\delta_i\right)$. The posterior probability for θ in a given bin (b_1, b_2, b_3, b_4) is then

$$\begin{aligned}
\mathbb{P}(\theta \in (b_1,b_2,b_3,b_4)|\{(t_i,y_i)\}_{i=1}^n) &\propto \int_{\theta_4^{\min}+b_4\cdot\delta_4}^{\theta_4^{\min}+(b_4+1)\delta_4} \cdots \int_{\theta_1^{\min}+b_1\cdot\delta_1}^{\theta_1^{\min}+(b_1+1)\delta_1} \pi(\theta)\,\mathcal{L}(\theta|\{(t_i,y_i)\}_{i=1}^n)\,d\theta_1\cdots d\theta_4 \\
&\approx (\delta_1\delta_2\delta_3\delta_4) \times \pi\left(\theta_1^{\min} + \left(b_1+\tfrac{1}{2}\right)\delta_1, \ldots, \theta_4^{\min} + \left(b_4+\tfrac{1}{2}\right)\delta_4\right) \\
&\qquad \times \mathcal{L}\left(\left(\theta_1^{\min} + \left(b_1+\tfrac{1}{2}\right)\delta_1, \ldots, \theta_4^{\min} + \left(b_4+\tfrac{1}{2}\right)\delta_4\right)\middle|\{(t_i,y_i)\}_{i=1}^n\right)
\end{aligned}$$

given that the bin sizes $\delta_1, \ldots, \delta_4$ are small enough, where $\mathcal{L}$ is the appropriate likelihood function for each dataset. BESTish thus computes the posterior probabilities across the $n_1 \times \cdots \times n_4$ bins, and combines them into the joint posterior distribution across the parameter space.

Contribution of clinical variates in selection rate heterogeneity

We study the impact of clinical information on the selection rates inferred for the same CH variant across different patients. Longitudinal biological measurements (e.g., in blood) are summarized for each individual with two statistics: mean value and change rate, where the latter is defined as the slope of the linear regression with respect to age.

We separate two sources of heterogeneity in CH fitness: one associated with specific genes and one comprised of other internal and external factors. Comparing the variance of BESTish-inferred CH selection rates and that of the fitness residuals (i.e., selection rates substracted by the gene-specific fitness average) reveals that gene identity explains 24.8% of the selection rate variance in our dataset (100 observations from 89 patients comprising 20 variants across 10 genes).

To analyze the power of clinical information to explain the remaining 75.2% of the fitness variance, we first encode categorical covariates numerically and exclude clinical features with $> 50\%$ missingness, resulting in a total of 109 clinical covariates. We split the data into a training set ($\approx 80\%$ of the observations) and validation set ($\approx 20\%$), such that data from the same patient appears in only one subsample. We train a random forest to predict the fitness residuals from patient-specific clinical covariates using the training set (500 trees; max features = 'sqrt'; min samples split = 5; min samples leaf = 2), where any missing data is approximated with k-nearest-neighbor imputation where $k = 5$. To evaluate the predictive value of the random forest, the same imputation strategy is then performed on the validation set, and model performance is evaluated using 5-fold GroupKFold cross-validation. However, out-of-sample performance for predicting gene-adjusted residuals is low (grouped cross-validation R-squared $\approx -0.09 \pm 0.14$), indicating limited out-of-sample predictive performance, likely due to the small sample size. Feature ranking was quantified using permutation importance with 30 repeats, computed as the mean decrease in R-squared after randomly permuting a single feature's values. However, due to the low predictive power, the feature rankings should be seen as hypothesis-generating rather than predictive. We further complement the random forest approach with SHAP (SHapley Additive exPlanations) values computed via TreeExplainer [Lundberg et al., 2020].

Despite the small sample size, both permutation importance (Figure 3D) and SHAP (Figure 3E) agree on the top-ranking features: change rates of ALT, CRP and ferritin (permutation importance mean $\pm$ standard deviation: 0.0386 ± 0.0072, 0.0284 ± 0.0067, and 0.0221 ± 0.0038, respectively). Spearman rank correlation further confirms that CH selection rates are negatively associated with the change rates of ALT and CRP with statistical significance ($\rho = -0.31$, p-value $= 0.0014$ and $\rho = -0.28$, p-value $= 0.0055$, respectively) (Figure 3F-G). In contrast, ferritin change rate is positively associated with increased selection rate but without significance ($\rho = 0.14$, p-value $= 0.1781$) (Figure 3H).

Code availability

BESTish is available at `https://github.com/dinhngockhanh/BESTish`.

Acknowledgments

RYW, KND and KT acknowledge the support from the Herbert and Florence Irving Institute for Cancer Dynamics at Columbia University. RYW, KYK and MK are supported by NIH grant P01CA265748. KYK is also supported by NIH grant R35 HL155672. GP is supported by NSF grants DMS 2216765 and CMMI 2452829.

Competing interests

The authors declare no competing interests.

References

D. F. Anderson and T. G. Kurtz. *Stochastic analysis of biochemical systems.* Springer, 2015.

T. Antal and P. Krapivsky. Exact solution of a two-type branching process: models of tumor progression. *Journal of Statistical Mechanics: Theory and Experiment*, 2011(08):P08018, 2011.

S. Avagyan and L. I. Zon. Clonal hematopoiesis and inflammation–the perpetual cycle. *Trends in cell biology*, 33 (8):695–707, 2023.

A. G. Bick, J. S. Weinstock, S. K. Nandakumar, C. P. Fulco, E. L. Bao, S. M. Zekavat, M. D. Szeto, X. Liao, M. J. Leventhal, J. Nasser, et al. Inherited causes of clonal haematopoiesis in 97,691 whole genomes. *Nature*, 586(7831):763–768, 2020.

P. Billingsley. *Convergence of probability measures.* John Wiley & Sons, 1999.

R. L. Bowman, L. Busque, and R. L. Levine. Clonal hematopoiesis and evolution to hematopoietic malignancies. *Cell Stem Cell*, 22(2):157–170, 2018.

K. Busch, K. Klapproth, M. Barile, M. Flossdorf, T. Holland-Letz, S. M. Schlenner, M. Reth, T. Höfer, and H.-R. Rodewald. Fundamental properties of unperturbed haematopoiesis from stem cells in vivo. *Nature*, 518(7540): 542–546, 2015.

S. N. Catlin, L. Busque, R. E. Gale, P. Guttorp, and J. L. Abkowitz. The replication rate of human hematopoietic stem cells in vivo. *Blood, The Journal of the American Society of Hematology*, 117(17):4460–4466, 2011.

D. Chakravarty, J. Gao, S. Phillips, R. Kundra, H. Zhang, J. Wang, J. E. Rudolph, R. Yaeger, T. Soumerai, M. H. Nissan, et al. Oncokb: a precision oncology knowledge base. *JCO Precision Oncology*, 1:1–16, 2017.

S. Chen, S. Benbarche, and O. Abdel-Wahab. Splicing factor mutations in hematologic malignancies. *Blood, The Journal of the American Society of Hematology*, 138(8):599–612, 2021.

G. S. Choudhary, A. Pellagatti, B. Agianian, M. A. Smith, T. D. Bhagat, S. Gordon-Mitchell, S. Sahu, S. Pandey, N. Shah, S. Aluri, et al. Activation of targetable inflammatory immune signaling is seen in myelodysplastic syndromes with SF3B1 mutations. *eLife*, 11:e78136, 2022.

C. C. Coombs, A. Zehir, S. M. Devlin, A. Kishtagari, A. Syed, P. Jonsson, D. M. Hyman, D. B. Solit, M. E. Robson, J. Baselga, et al. Therapy-related clonal hematopoiesis in patients with non-hematologic cancers is common and associated with adverse clinical outcomes. *Cell Stem Cell*, 21(3):374–382, 2017.

J. Cosgrove, L. S. Hustin, R. J. de Boer, and L. Perié. Hematopoiesis in numbers. *Trends in Immunology*, 42(12): 1100–1112, 2021.

S. C. Dentro, D. C. Wedge, and P. Van Loo. Principles of reconstructing the subclonal architecture of cancers. *Cold Spring Harbor Perspectives in Medicine*, 7(8):a026625, 2017.

R. Durrett. Branching Process Models of Cancer. Springer, 2015.

R. Durrett and S. Moseley. Evolution of resistance and progression to disease during clonal expansion of cancer. *Theoretical Population Biology*, 77(1):42–48, 2010.

R. Durrett, J. Foo, K. Leder, J. Mayberry, and F. Michor. Intratumor heterogeneity in evolutionary models of tumor progression. *Genetics*, 188(2):461–477, 2011.

S. N. Ethier and T. G. Kurtz. *Markov processes: Characterization and Convergence*. John Wiley & Sons, 2009.

M. A. Fabre, J. G. de Almeida, E. Fiorillo, E. Mitchell, A. Damaskou, J. Rak, V. Orrù, M. Marongiu, M. S. Chapman, M. Vijayabaskar, et al. The longitudinal dynamics and natural history of clonal haematopoiesis. *Nature*, 606(7913):335–342, 2022.

M. A. Florez, B. T. Tran, T. K. Wathan, J. DeGregori, E. M. Pietras, and K. Y. King. Clonal hematopoiesis: Mutation-specific adaptation to environmental change. *Cell Stem Cell*, 29(6):882–904, 2022.

P. Getto, A. Marciniak-Czochra, Y. Nakata, et al. Global dynamics of two-compartment models for cell production systems with regulatory mechanisms. *Mathematical Biosciences*, 245(2):258–268, 2013.

D. Hormaechea-Agulla, K. A. Matatall, D. T. Le, B. Kain, X. Long, P. Kus, R. Jaksik, G. A. Challen, M. Kimmel, and K. Y. King. Chronic infection drives dnmt3a-loss-of-function clonal hematopoiesis via ifnγ signaling. *Cell Stem Cell*, 28(8):1428–1442, 2021.

B. Johnson, Y. Shuai, J. Schweinsberg, and K. Curtius. clonerate: fast estimation of single-cell clonal dynamics using coalescent theory. *Bioinformatics*, 39(9):btad561, 2023.

S. P. Kar, P. M. Quiros, M. Gu, T. Jiang, J. Mitchell, R. Langdon, V. Iyer, C. Barcena, M. Vijayabaskar, M. A. Fabre, et al. Genome-wide analyses of 200,453 individuals yield new insights into the causes and consequences of clonal hematopoiesis. *Nature Genetics*, 54(8):1155–1166, 2022.

I. Karatzas and S. Shreve. *Brownian motion and stochastic calculus*. Springer, 2014.

N. L. Komarova, C. Rignot, A. G. Fleischman, and D. Wodarz. Dynamically adjusted cell fate decisions and resilience to mutant invasion during steady-state hematopoiesis revealed by an experimentally parameterized mathematical model. *Proceedings of the National Academy of Sciences*, 121(38):e2321525121, 2024.

H. Lee-Six, N. F. Øbro, M. S. Shepherd, S. Grossmann, K. Dawson, M. Belmonte, R. J. Osborne, B. J. Huntly, I. Martincorena, E. Anderson, et al. Population dynamics of normal human blood inferred from somatic mutations. *Nature*, 561(7724):473–478, 2018.

S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee. From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2(1):2522–5839, 2020.

T. McKerrell, N. Park, T. Moreno, C. S. Grove, H. Ponstingl, J. Stephens, C. Crawley, J. Craig, M. A. Scott, C. Hodkinson, et al. Leukemia-associated somatic mutations drive distinct patterns of age-related clonal hemopoiesis. *Cell Reports*, 10(8):1239–1245, 2015.

S. Meleard and V. Bansaye. *Stochastic Models for Structured Populations: Scaling Limits and Long Time Behavior*, volume 1. Springer, 2015.

E. Mitchell, M. Spencer Chapman, N. Williams, K. J. Dawson, N. Mende, E. F. Calderbank, H. Jung, T. Mitchell, T. H. Coorens, D. H. Spencer, et al. Clonal dynamics of haematopoiesis across the human lifespan. *Nature*, 606(7913):343–350, 2022.

M. D. Nicholson and T. Antal. Competing evolutionary paths in growing populations with applications to multidrug resistance. *PLOS Computational Biology*, 15(4):e1006866, 2019.

N. A. Robertson, E. Latorre-Crespo, M. Terradas-Terradas, J. Lemos-Portela, A. C. Purcell, B. J. Livesey, R. F. Hillary, L. Murphy, A. Fawkes, L. MacGillivray, et al. Longitudinal dynamics of clonal hematopoiesis identifies gene-specific fitness effects. *Nature Medicine*, 28(7):1439–1446, 2022.

S. A. Sisson, Y. Fan, and M. Beaumont. *Handbook of Approximate Bayesian Computation*. CRC Press, 2018.

R.-Y. Wang, M. Kimmel, and G. Pang. Stochastic dynamics of two-compartment cell proliferation models with regulatory mechanisms for hematopoiesis. *Journal of Mathematical Biology*, 91(2):18, 2025.

C. J. Watson, A. Papula, G. Y. Poon, W. H. Wong, A. L. Young, T. E. Druley, D. S. Fisher, and J. R. Blundell. The evolutionary dynamics and fitness landscape of clonal hematopoiesis. *Science*, 367(6485):1449–1454, 2020.

W. Whitt. *Stochastic-process limits: an introduction to stochastic-process limits and their application to queues*. Springer, 2002.

S. Winter, K. S. Götze, J. S. Hecker, K. H. Metzeler, B. Guezguez, K. Woods, H. Medyouf, A. Schäffer, M. Schmitz, R. Wehner, et al. Clonal hematopoiesis and its impact on the aging osteo-hematopoietic niche. *Leukemia*, 38 (5):936–946, 2024.

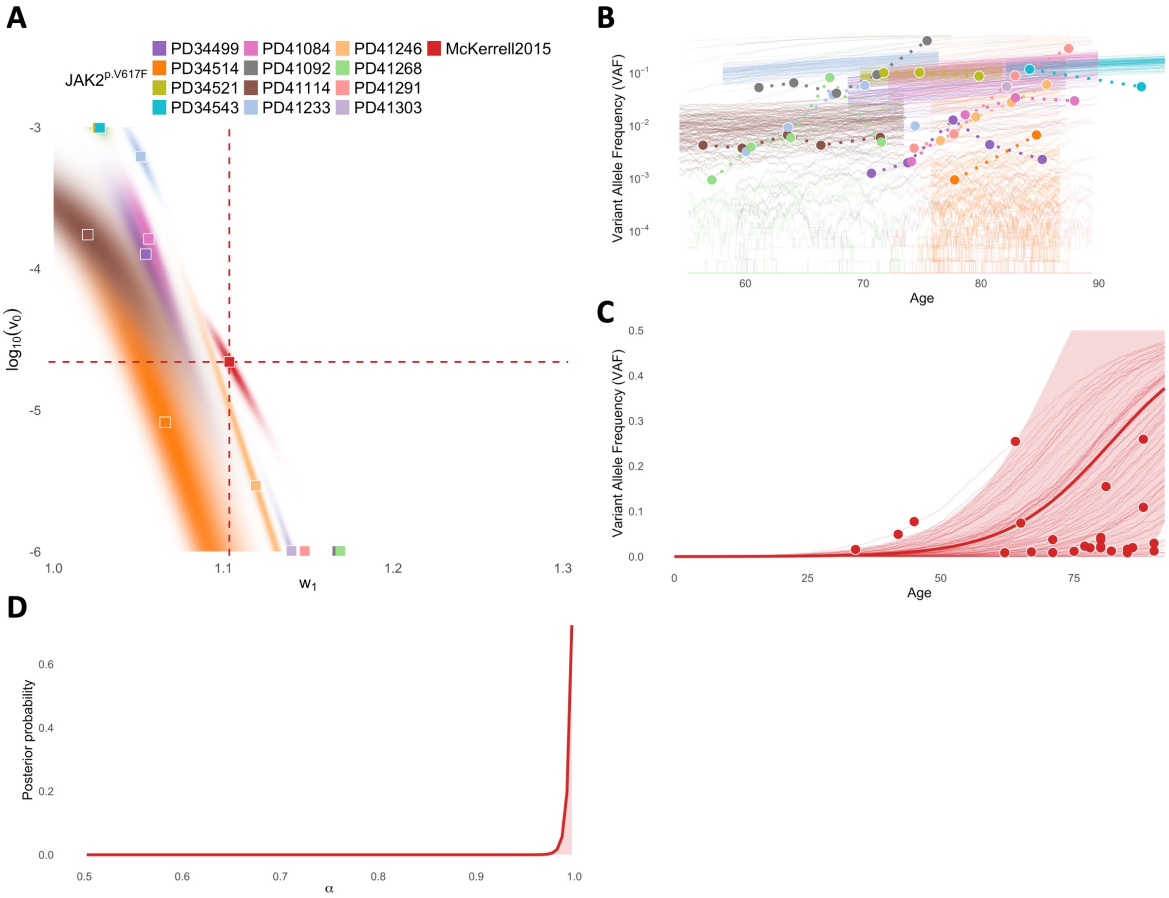
D. Wodarz and N. Komarova. *Dynamics of cancer: mathematical foundations of oncology*. World Scientific, 2014.

W. J. Wong, C. Emdin, A. G. Bick, S. M. Zekavat, A. Niroula, J. P. Pirruccello, L. Dichtel, G. Griffin, M. M. Uddin, C. J. Gibson, et al. Clonal haematopoiesis and risk of chronic liver disease. *Nature*, 616(7958):747–754, 2023.

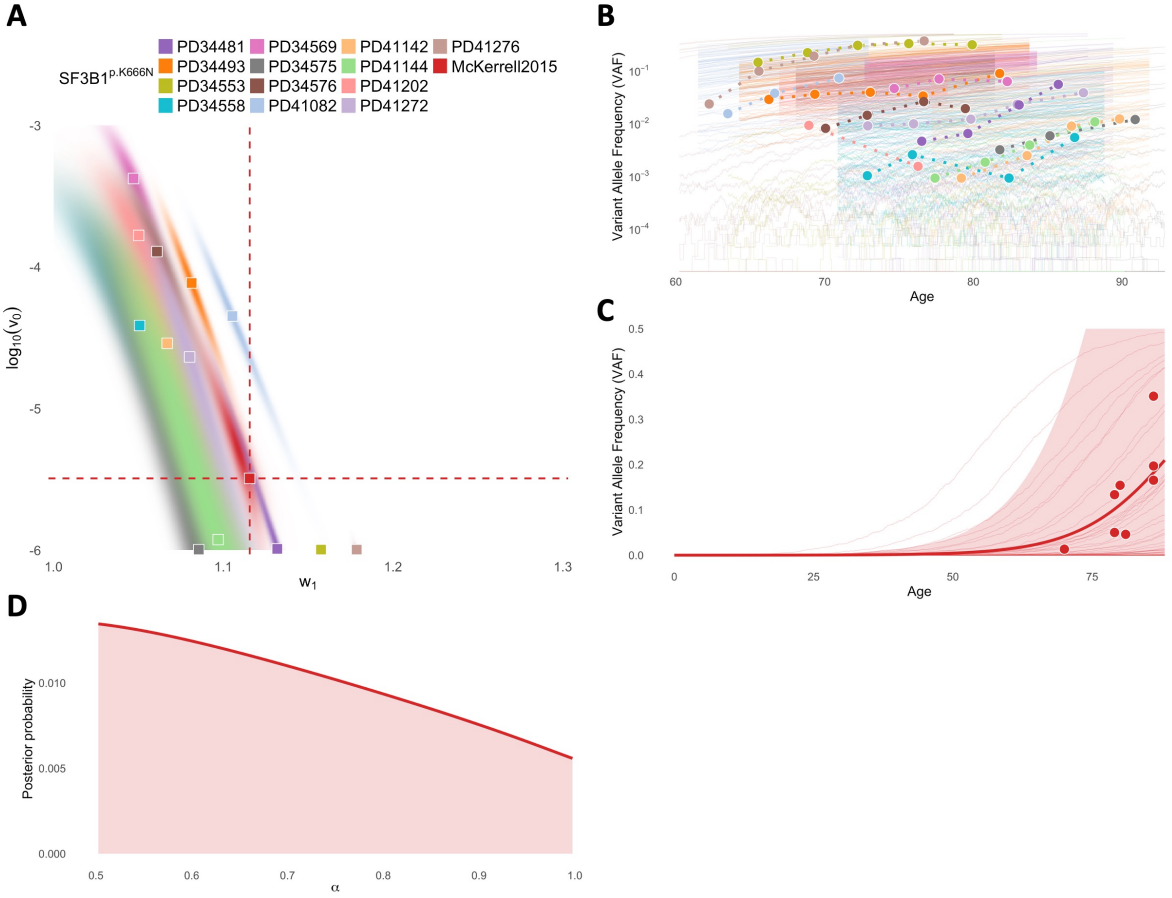
H. Zhang, B. Wilmot, D. Bottomly, K.-H. T. Dao, E. Stevens, C. A. Eide, V. Khanna, A. Rofelty, S. Savage, A. Reister Schultz, et al. Genomic landscape of neutrophilic leukemias of ambiguous diagnosis. *Blood, The Journal of the American Society of Hematology*, 134(11):867–879, 2019.

R. Zhang and I. Bozic. Accumulation of oncogenic mutations during progression from healthy tissue to cancer. *Bulletin of Mathematical Biology*, 86(12):1–33, 2024.

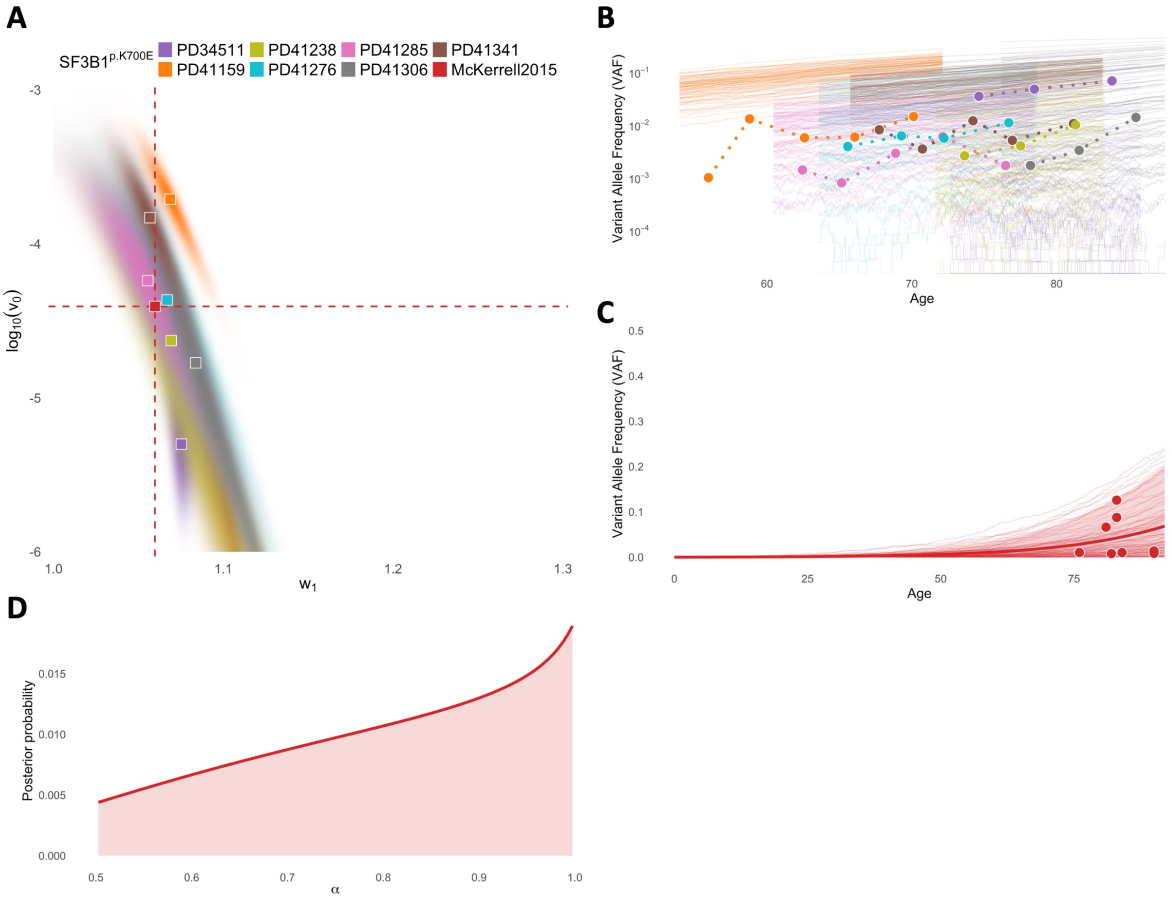
S. Zhang, X. Xiao, Y. Yi, X. Wang, L. Zhu, Y. Shen, D. Lin, and C. Wu. Tumor initiation and early tumorigenesis: molecular mechanisms and interventional targets. *Signal Transduction and Targeted Therapy*, 9(1):149, 2024.



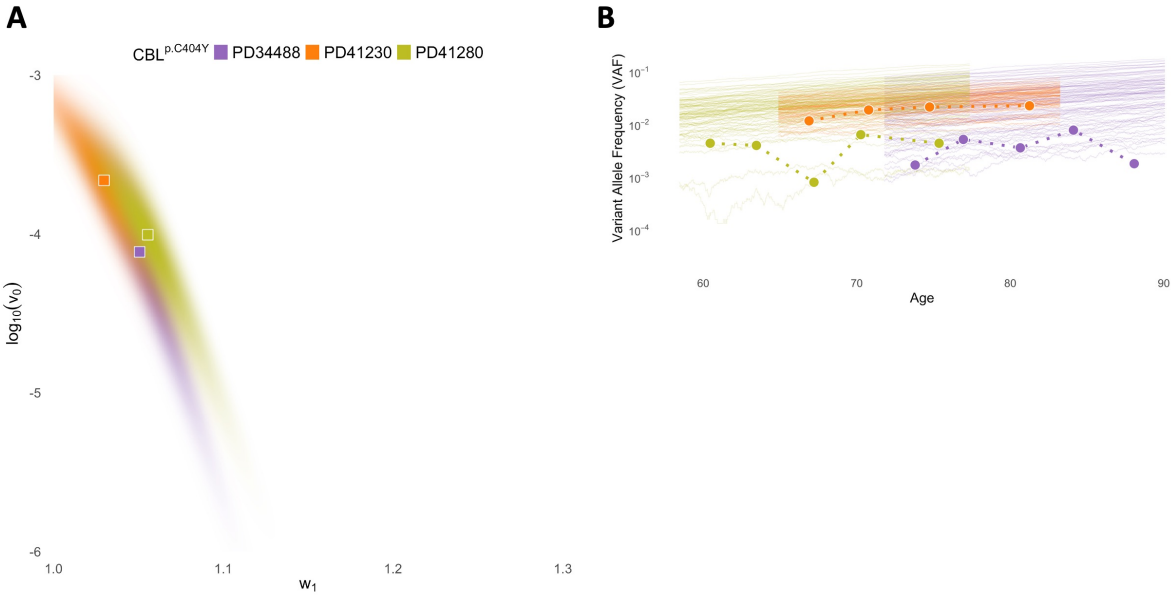
Supplementary Figure 1: Inference of *JAK2-V617F*'s mutation rate and selection rate from cohort samples in [McKerrell et al., 2015] ("McKerrell2015") and time-series data in [Fabre et al., 2022]. **A**: Joint posterior distributions for $\log_{10}(v_0)$ and w_1. Squares = MAP estimates. **B**: 100 simulated VAF trajectories in logscale assuming MAP estimates (thin lines), against observed VAFs (circles, connected by dashed lines) for samples in [Fabre et al., 2022]. **C**: Expected values (thick lines) and predicted 95%CI regions (shaded areas) for VAF trajectories assuming MAP estimates, against observed VAFs (circles) and simulations with MAP parameters (thin lines) for samples in [McKerrell et al., 2015]. Colors in **B-D** correspond to datasets in **A**.



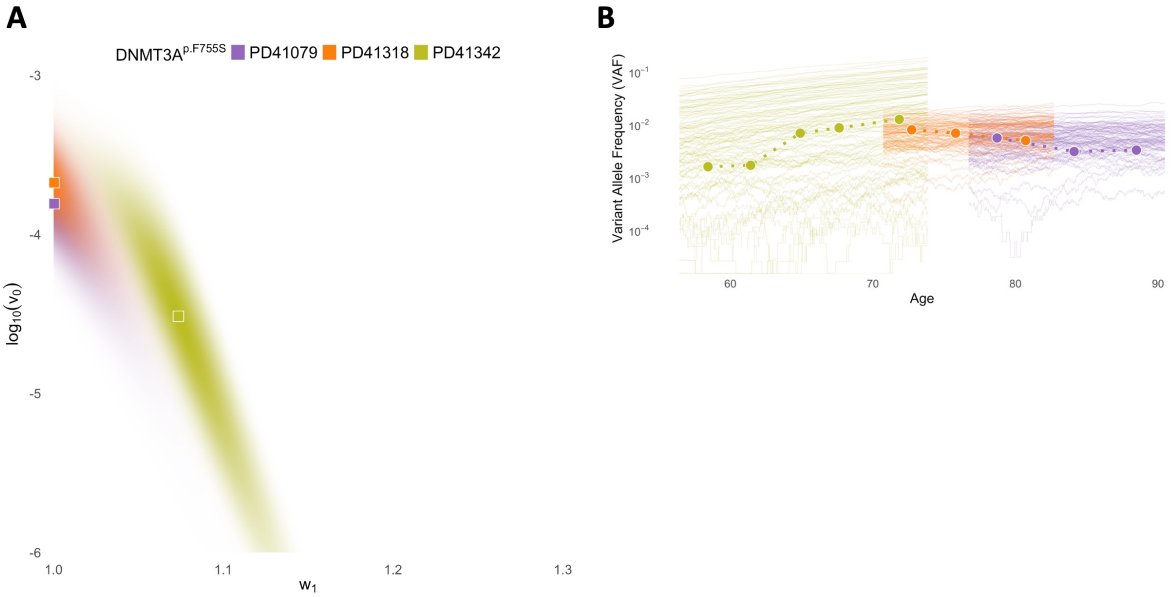
Supplementary Figure 2: Inference of *SF3B1-K666N*'s mutation rate and selection rate from cohort samples in [McKerrell et al., 2015] ("McKerrell2015") and time-series data in [Fabre et al., 2022]. **A**: Joint posterior distributions for $\log_{10}(v_0)$ and w_1. Squares = MAP estimates. **B**: 100 simulated VAF trajectories in logscale assuming MAP estimates (thin lines), against observed VAFs (circles, connected by dashed lines) for samples in [Fabre et al., 2022]. **C**: Expected values (thick lines) and predicted 95%CI regions (shaded areas) for VAF trajectories assuming MAP estimates, against observed VAFs (circles) and simulations with MAP parameters (thin lines) for samples in [McKerrell et al., 2015]. Colors in **B-D** correspond to datasets in **A**.



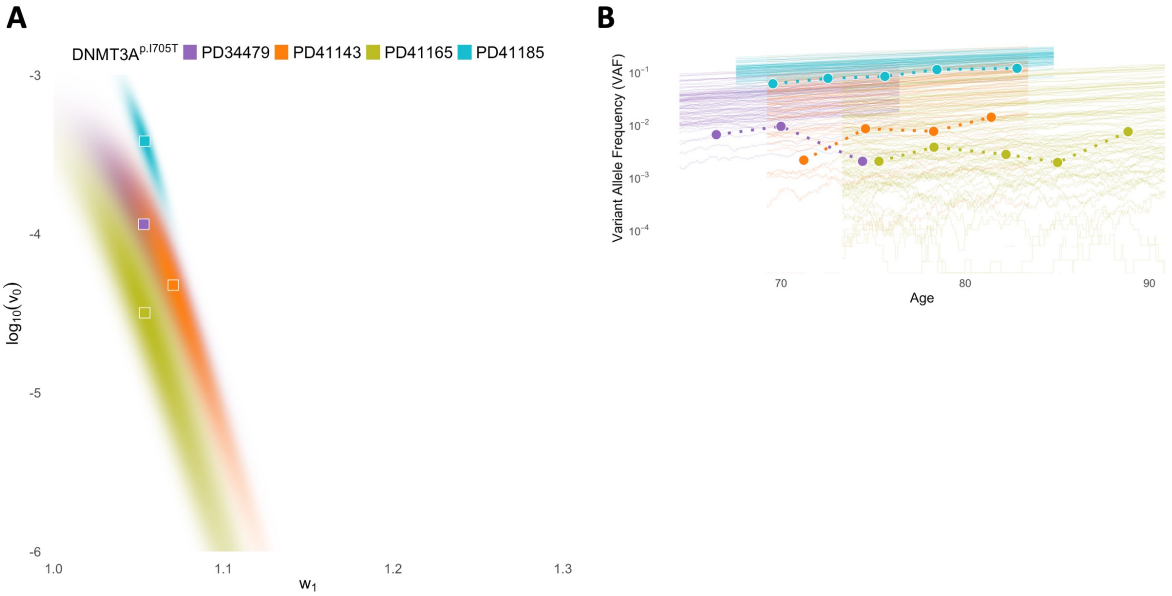
Supplementary Figure 3: Inference of *SF3B1-K700E*'s mutation rate and selection rate from cohort samples in [McKerrell et al., 2015] ("McKerrell2015") and time-series data in [Fabre et al., 2022]. **A**: Joint posterior distributions for $\log_{10}(v_0)$ and w_1. Squares = MAP estimates. **B**: 100 simulated VAF trajectories in logscale assuming MAP estimates (thin lines), against observed VAFs (circles, connected by dashed lines) for samples in [Fabre et al., 2022]. **C**: Expected values (thick lines) and predicted 95%CI regions (shaded areas) for VAF trajectories assuming MAP estimates, against observed VAFs (circles) and simulations with MAP parameters (thin lines) for samples in [McKerrell et al., 2015]. Colors in **B-D** correspond to datasets in **A**.



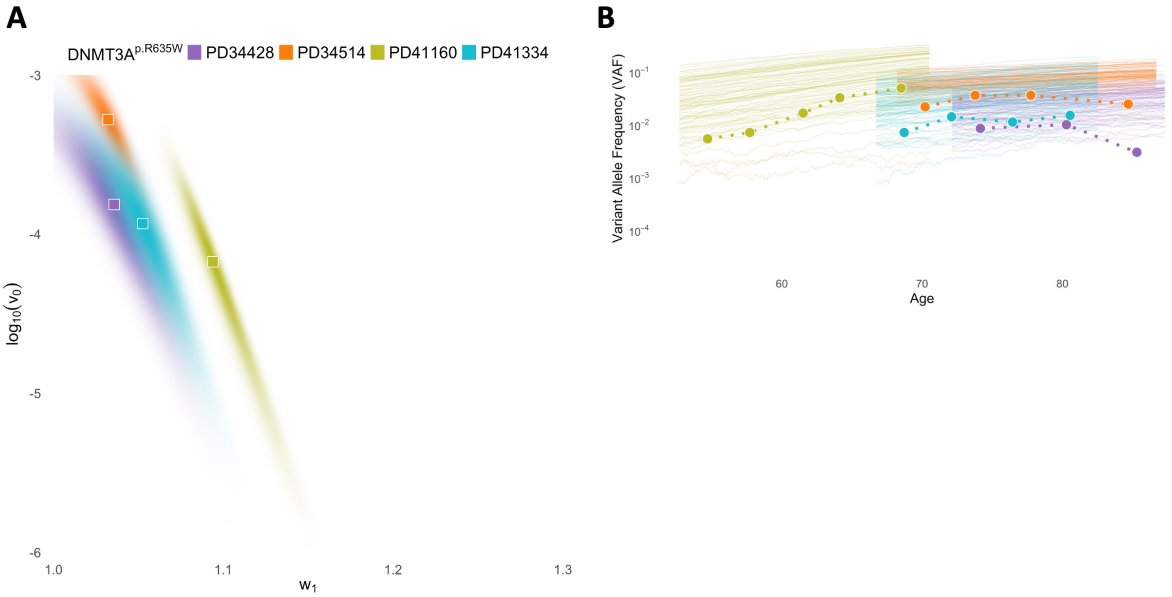
Supplementary Figure 4: Inference of *CBL-C404Y*'s mutation rate and selection rate from time-series data in [Fabre et al., 2022]. **A**: Joint posterior distributions for $\log_{10}(v_0)$ and w_1. Squares = MAP estimates. **B**: 100 simulated VAF trajectories in logscale assuming MAP estimates (thin lines), against observed VAFs (circles, connected by dashed lines). Colors in **B** correspond to datasets in **A**.



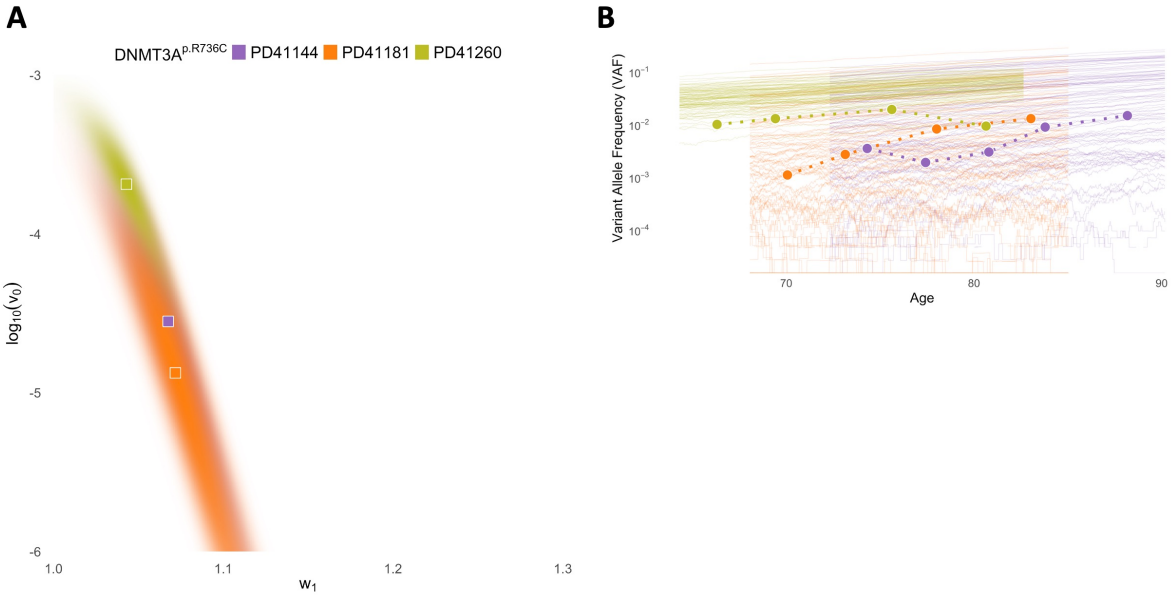
Supplementary Figure 5: Inference of *DNMT3A-F755S*'s mutation rate and selection rate from time-series data in [Fabre et al., 2022]. **A**: Joint posterior distributions for $\log_{10}(v_0)$ and w_1. Squares = MAP estimates. **B**: 100 simulated VAF trajectories in logscale assuming MAP estimates (thin lines), against observed VAFs (circles, connected by dashed lines). Colors in **B** correspond to datasets in **A**.



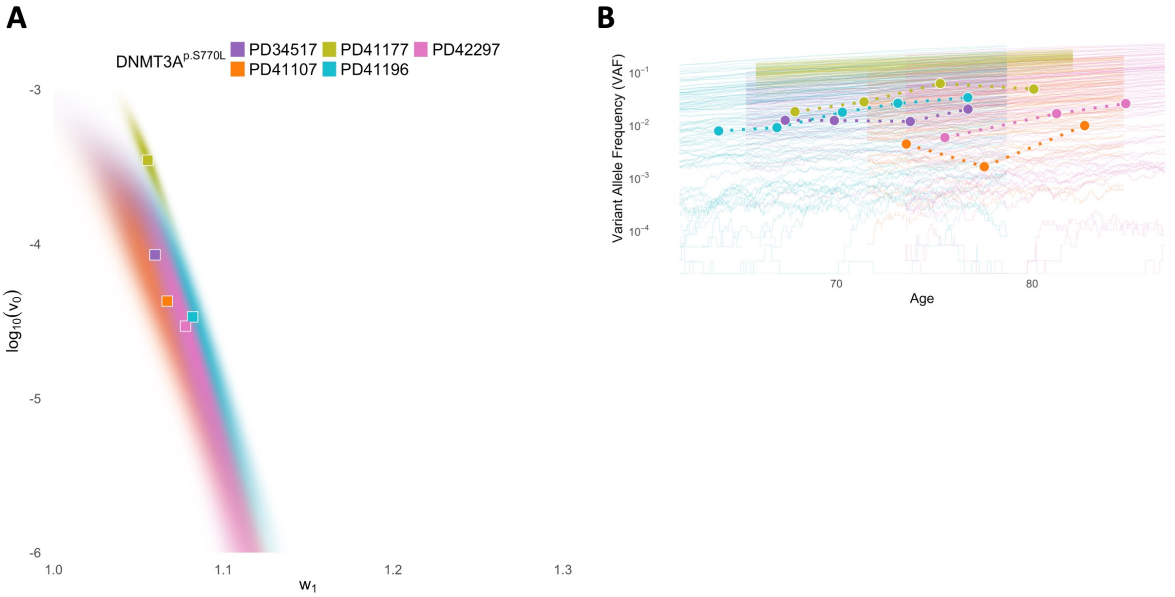
Supplementary Figure 6: Inference of *DNMT3A-I705T*'s mutation rate and selection rate from time-series data in [Fabre et al., 2022]. **A**: Joint posterior distributions for $\log_{10}(v_0)$ and w_1. Squares = MAP estimates. **B**: 100 simulated VAF trajectories in logscale assuming MAP estimates (thin lines), against observed VAFs (circles, connected by dashed lines). Colors in **B** correspond to datasets in **A**.



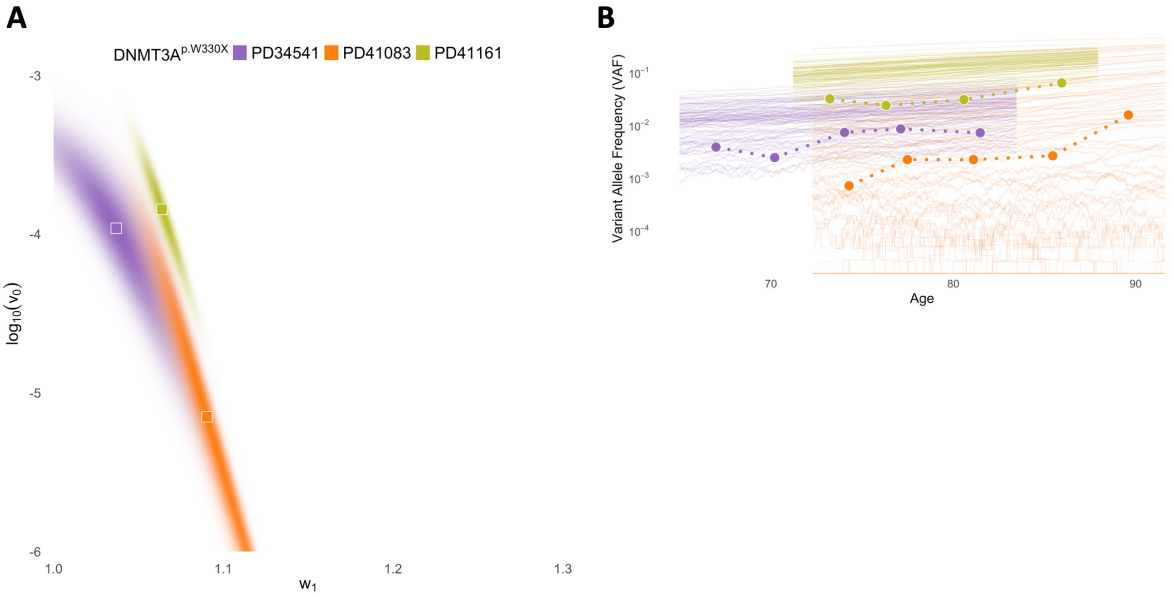
Supplementary Figure 7: Inference of *DNMT3A-R635W*'s mutation rate and selection rate from time-series data in [Fabre et al., 2022]. **A**: Joint posterior distributions for $\log_{10}(v_0)$ and w_1. Squares = MAP estimates. **B**: 100 simulated VAF trajectories in logscale assuming MAP estimates (thin lines), against observed VAFs (circles, connected by dashed lines). Colors in **B** correspond to datasets in **A**.



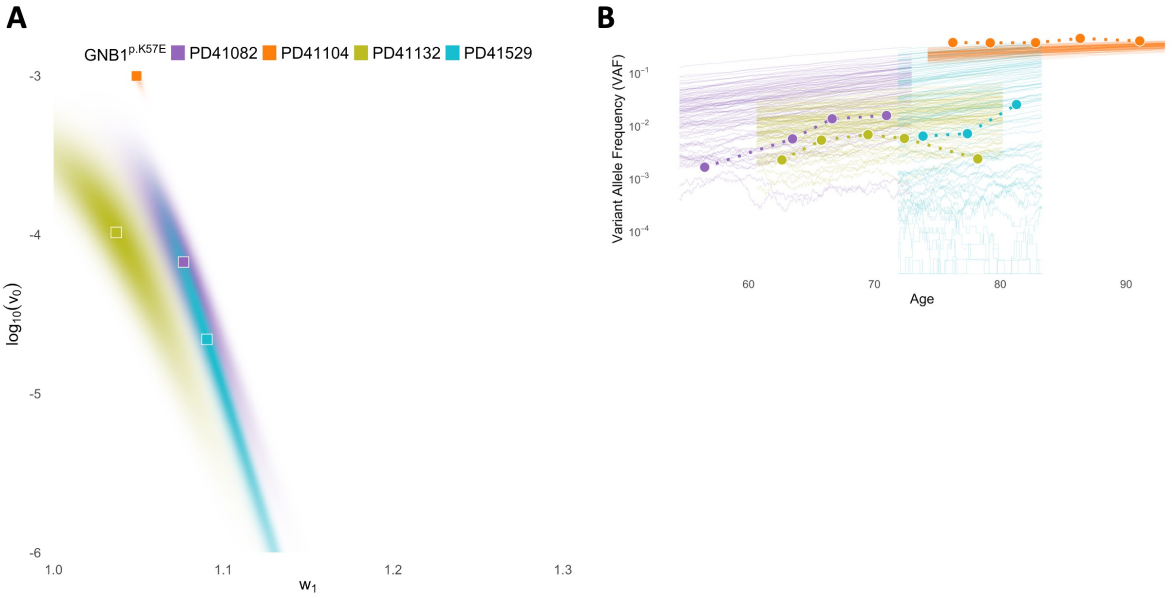
Supplementary Figure 8: Inference of *DNMT3A-R736C*'s mutation rate and selection rate from time-series data in [Fabre et al., 2022]. **A**: Joint posterior distributions for $\log_{10}(v_0)$ and w_1. Squares = MAP estimates. **B**: 100 simulated VAF trajectories in logscale assuming MAP estimates (thin lines), against observed VAFs (circles, connected by dashed lines). Colors in **B** correspond to datasets in **A**.



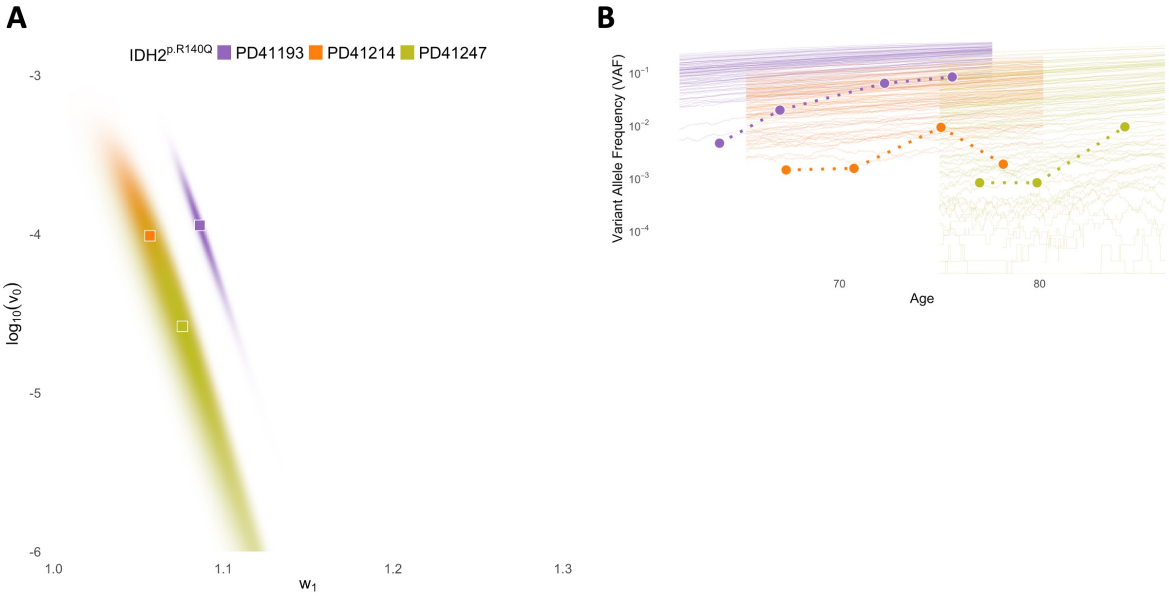
Supplementary Figure 9: Inference of *DNMT3A-S770L*'s mutation rate and selection rate from time-series data in [Fabre et al., 2022]. **A**: Joint posterior distributions for $\log_{10}(v_0)$ and w_1. Squares = MAP estimates. **B**: 100 simulated VAF trajectories in logscale assuming MAP estimates (thin lines), against observed VAFs (circles, connected by dashed lines). Colors in **B** correspond to datasets in **A**.



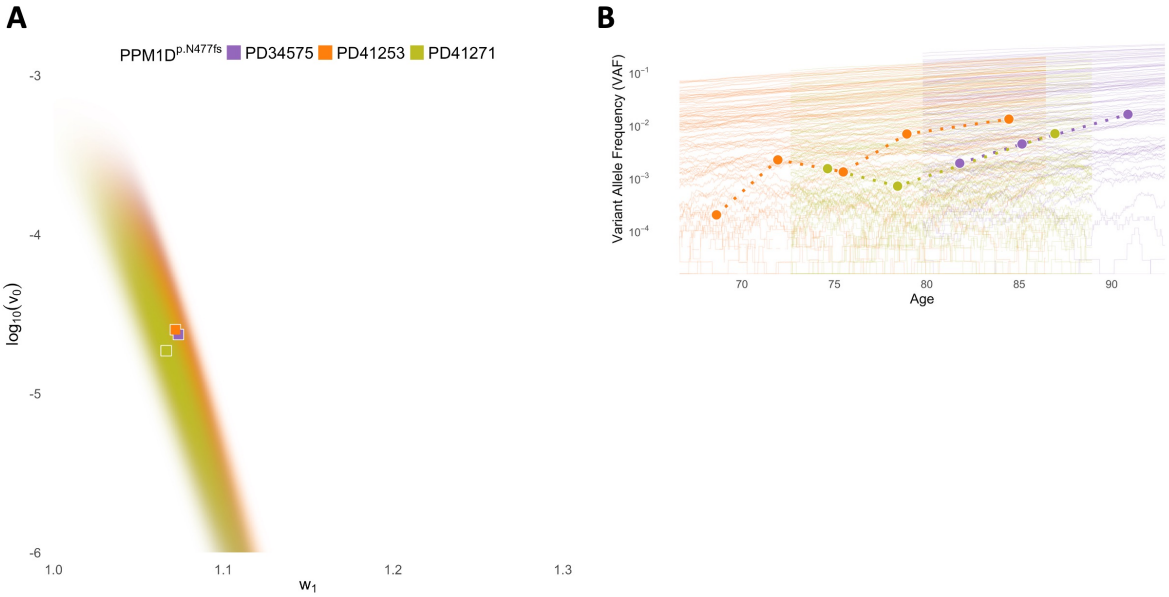
Supplementary Figure 10: Inference of *DNMT3A-W330X*'s mutation rate and selection rate from time-series data in [Fabre et al., 2022]. **A**: Joint posterior distributions for $\log_{10}(v_0)$ and w_1. Squares = MAP estimates. **B**: 100 simulated VAF trajectories in logscale assuming MAP estimates (thin lines), against observed VAFs (circles, connected by dashed lines). Colors in **B** correspond to datasets in **A**.



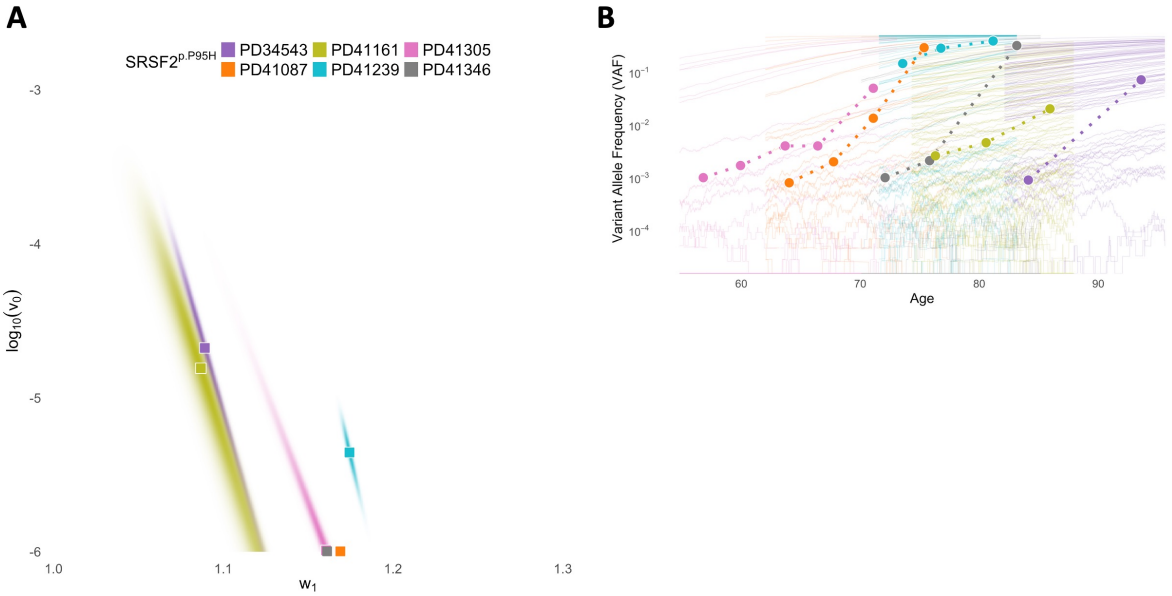
Supplementary Figure 11: Inference of *GNB1-K57E*'s mutation rate and selection rate from time-series data in [Fabre et al., 2022]. **A**: Joint posterior distributions for $\log_{10}(v_0)$ and w_1. Squares = MAP estimates. **B**: 100 simulated VAF trajectories in logscale assuming MAP estimates (thin lines), against observed VAFs (circles, connected by dashed lines). Colors in **B** correspond to datasets in **A**.



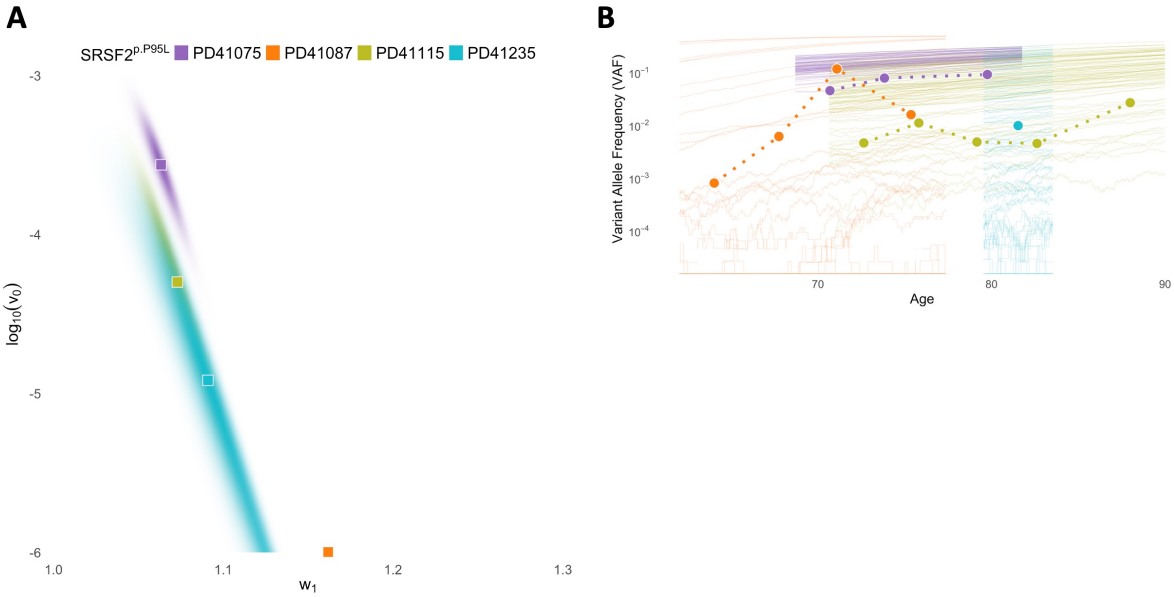
Supplementary Figure 12: Inference of *IDH2-R140Q*'s mutation rate and selection rate from time-series data in [Fabre et al., 2022]. **A**: Joint posterior distributions for $\log_{10}(v_0)$ and w_1. Squares = MAP estimates. **B**: 100 simulated VAF trajectories in logscale assuming MAP estimates (thin lines), against observed VAFs (circles, connected by dashed lines). Colors in **B** correspond to datasets in **A**.



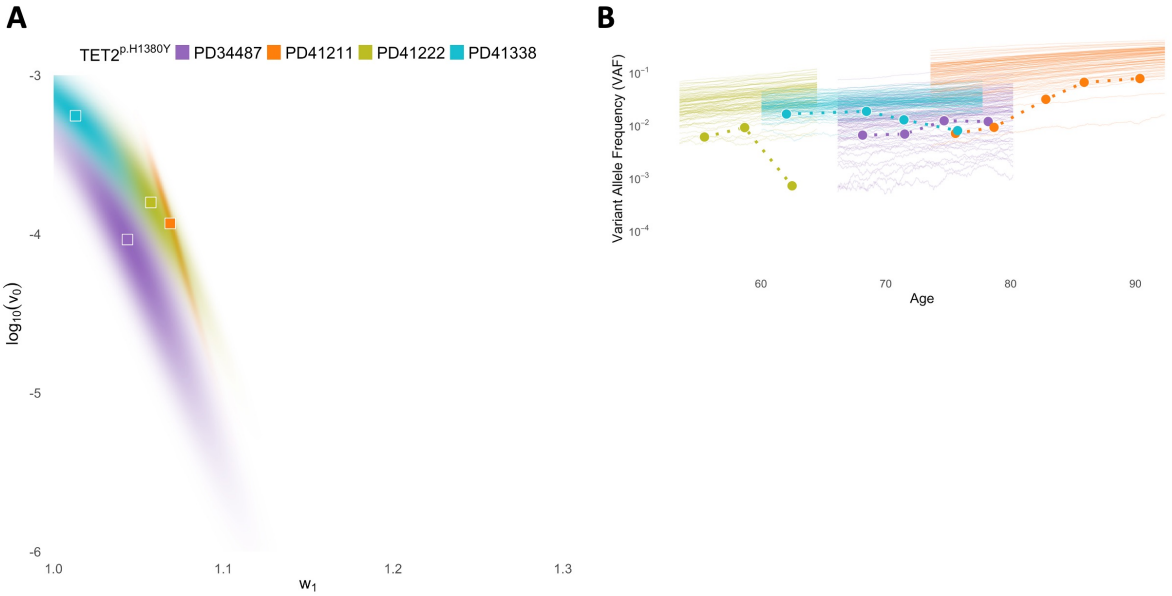
Supplementary Figure 13: Inference of *PPM1D-N477fs*'s mutation rate and selection rate from time-series data in [Fabre et al., 2022]. **A**: Joint posterior distributions for $\log_{10}(v_0)$ and w_1. Squares = MAP estimates. **B**: 100 simulated VAF trajectories in logscale assuming MAP estimates (thin lines), against observed VAFs (circles, connected by dashed lines). Colors in **B** correspond to datasets in **A**.



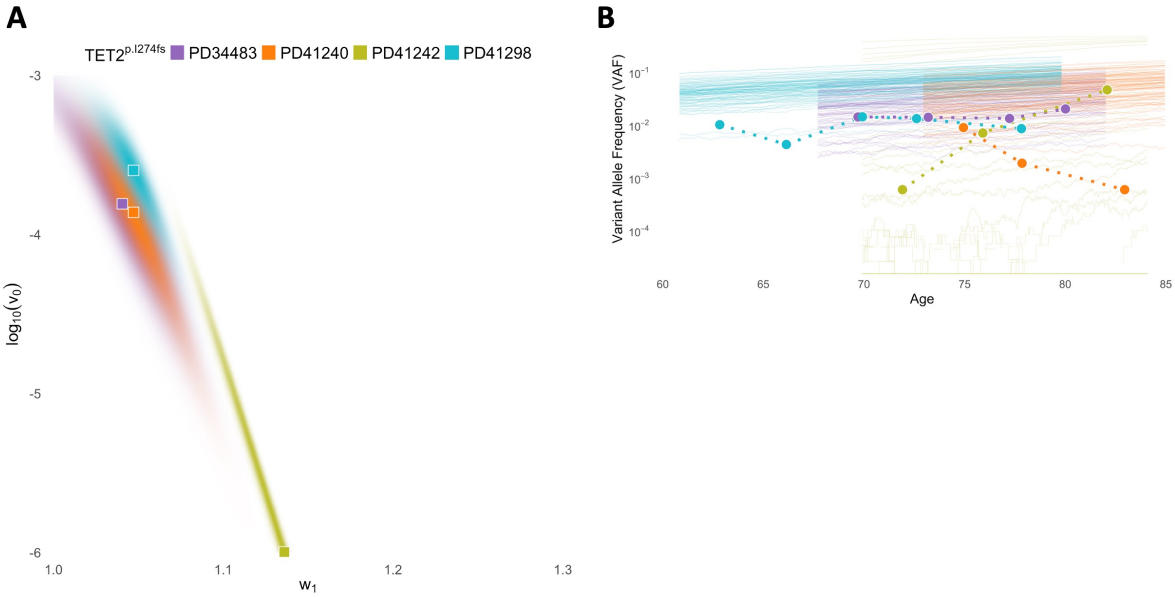
Supplementary Figure 14: Inference of *SRSF2-P95H*'s mutation rate and selection rate from time-series data in [Fabre et al., 2022]. **A**: Joint posterior distributions for $\log_{10}(v_0)$ and w_1. Squares = MAP estimates. **B**: 100 simulated VAF trajectories in logscale assuming MAP estimates (thin lines), against observed VAFs (circles, connected by dashed lines). Colors in **B** correspond to datasets in **A**.



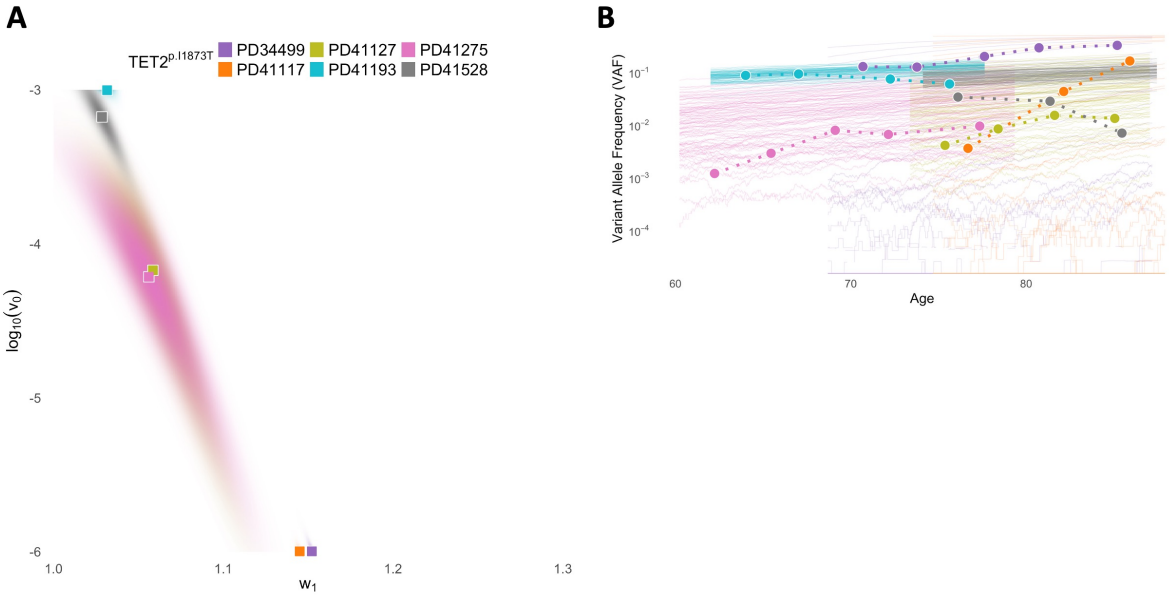
Supplementary Figure 15: Inference of *SRSF2-P95L*'s mutation rate and selection rate from time-series data in [Fabre et al., 2022]. **A**: Joint posterior distributions for $\log_{10}(v_0)$ and w_1. Squares = MAP estimates. **B**: 100 simulated VAF trajectories in logscale assuming MAP estimates (thin lines), against observed VAFs (circles, connected by dashed lines). Colors in **B** correspond to datasets in **A**.



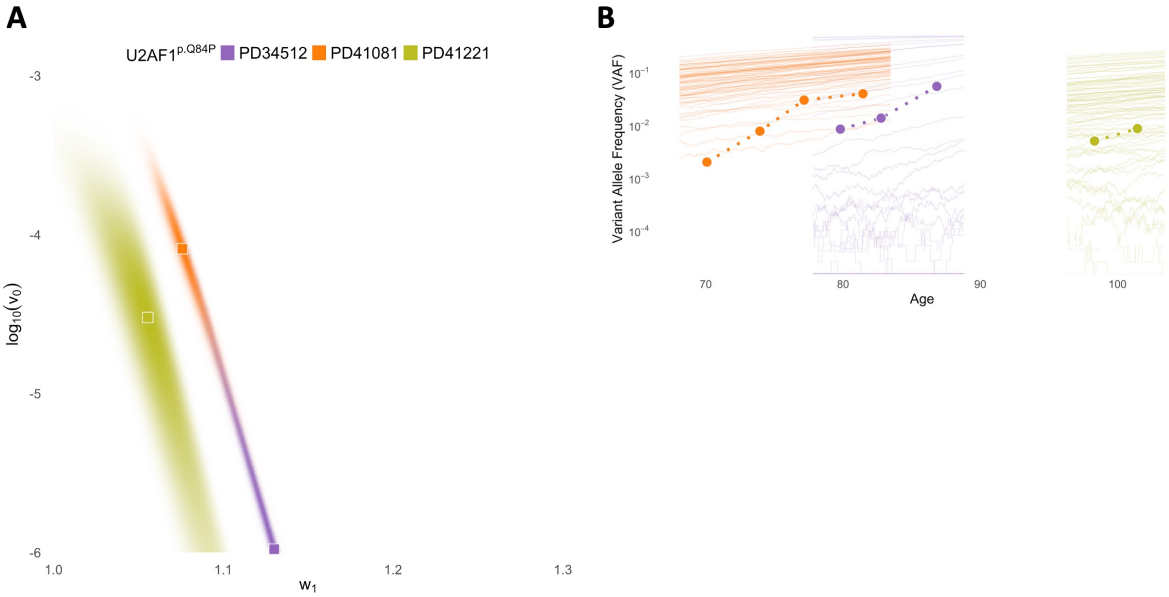
Supplementary Figure 16: Inference of *TET2-H1380Y*'s mutation rate and selection rate from time-series data in [Fabre et al., 2022]. **A**: Joint posterior distributions for $\log_{10}(v_0)$ and w_1. Squares = MAP estimates. **B**: 100 simulated VAF trajectories in logscale assuming MAP estimates (thin lines), against observed VAFs (circles, connected by dashed lines). Colors in **B** correspond to datasets in **A**.



Supplementary Figure 17: Inference of *TET2-I274fs*'s mutation rate and selection rate from time-series data in [Fabre et al., 2022]. **A**: Joint posterior distributions for $\log_{10}(v_0)$ and w_1. Squares = MAP estimates. **B**: 100 simulated VAF trajectories in logscale assuming MAP estimates (thin lines), against observed VAFs (circles, connected by dashed lines). Colors in **B** correspond to datasets in **A**.



Supplementary Figure 18: Inference of *TET2-I1873T*'s mutation rate and selection rate from time-series data in [Fabre et al., 2022]. **A**: Joint posterior distributions for $\log_{10}(v_0)$ and w_1. Squares = MAP estimates. **B**: 100 simulated VAF trajectories in logscale assuming MAP estimates (thin lines), against observed VAFs (circles, connected by dashed lines). Colors in **B** correspond to datasets in **A**.



Supplementary Figure 19: Inference of *U2AF1-Q84P*'s mutation rate and selection rate from time-series data in [Fabre et al., 2022]. **A**: Joint posterior distributions for $\log_{10}(v_0)$ and w_1. Squares = MAP estimates. **B**: 100 simulated VAF trajectories in logscale assuming MAP estimates (thin lines), against observed VAFs (circles, connected by dashed lines). Colors in **B** correspond to datasets in **A**.

A Appendix

A.1 Construction of Dynamics

Let $N_0^{(r)}(t)$ and $N_1^{(r)}(t)$ denote the number of WT (type 0) and CH (type 1) cells at time t with initial population consisting of r WT cells. We denote events in the dynamics by (j, a), where $j \in \{0, 1\}$ and $a \in \{b, m, d\}$ (birth, mutation, and death). Event (j, a) refers to a type j individual taking action a. We construct the model following the Poisson process formulation in [Ethier and Kurtz, 2009]. Let $\{P_{(j,a)}^{(r)} \mid r \in \mathbb{N}, 0 \leq j \leq n, a \in \{b, m, d\}\}$ be a set of independent rate 1 Poisson processes.

$$
\begin{aligned}
N_0^{(r)}(t) &= r + P_{(0,b)}^{(r)}\left(\int_0^t \lambda_0 N_0^{(r)}(s)ds\right) - P_{(0,m)}^{(r)}\left(\int_0^t v_0 N_0^{(r)}(s)ds\right) \\
&\quad - P_{(0,d)}^{(r)}\left(\int_0^t \frac{\alpha N_0^{(r)}(s)}{\sum_{j=0}^n N_j^{(r)}(s)} \sum_{j=0}^1 \lambda_j N_j^{(r)}(s)ds\right); \\
N_1^{(r)}(t) &= P_{(1,b)}^{(r)}\left(\int_0^t \lambda_1 N_1^{(r)}(s)ds\right) + P_{(0,m)}^{(r)}\left(\int_0^t v_0 N_0^{(r)}(s)ds\right) \\
&\quad - P_{(1,d)}^{(r)}\left(\int_0^t \frac{\alpha N_1^{(r)}(s)}{\sum_{j=0}^n N_j^{(r)}(s)} \sum_{j=0}^1 \lambda_j N_j^{(r)}(s)ds\right).
\end{aligned}
$$

For all $0 \leq j \leq 1$, division rate $\lambda_j > 0$ and mutation rate $v_0 > 0$. Since type 1 individuals cannot further mutate, $v_1 = 0$. For our model, we define fitness of a type j individual by $w_j := \lambda_j - v_j$.

B FLLN and FCLT

Proposition 1. *As $r \to \infty$, $|\overline{\mathbf{N}}^{(r)} - \overline{\mathbf{N}}| \to 0$ almost surely in Skorokhod space $(\mathbb{D}([0, \infty)), d_\infty)$, where $\overline{\mathbf{N}}$ is characterized by the following autonomous system of ODEs:*

$$
\begin{aligned}
&\overline{N}_0'(t) = w_0 \overline{N}_0(t) - \alpha \frac{\overline{N}_0(t)}{\sum_{j=0}^1 \overline{N}_j(t)} \sum_{j=0}^1 \lambda_j \overline{N}_j(t); \\
&\overline{N}_1'(t) = w_1 \overline{N}_1(t) + v_0 \overline{N}_0(t) - \alpha \frac{\overline{N}_1(t)}{\sum_{j=0}^1 \overline{N}_j(t)} \sum_{j=0}^1 \lambda_j \overline{N}_j(t), \\
&\textit{with initial condition } \overline{\mathbf{N}}(0) = (1, 0)^\top.
\end{aligned}
$$

Proof. We verify conditions of Theorem 2.1 in Ethier and Kurtz [2009]. Rewrite the system of differential equations as

$$
\overline{\mathbf{N}}' = \sum_{(j,a)} l_{(j,a)} \beta_{(j,a)}(\overline{\mathbf{N}}) := \mathbf{F}(\overline{\mathbf{N}}),
$$

where $l_{(j,a)}$ is a vector representing the change of population and $\beta_{(j,a)}$ is the rate for that event. For events associated with type 0 indiduals,

$$
\begin{aligned}
&l_{(0,b)} = (1, 0)^\top; l_{(0,m)} = (-1, 1)^\top; l_{(0,d)} = (-1, 0)^\top \text{ and} \\
&\beta_{(0,b)}(\overline{\mathbf{N}}) = \lambda_0 \overline{N}_0; \beta_{(0,m)}(\overline{\mathbf{N}}) = v_0 \overline{N}_0; \beta_{(0,d)}(\overline{\mathbf{N}}) = \frac{\alpha \overline{N}_0}{\sum_{j=0}^1 \overline{N}_j} \sum_{j=0}^1 \lambda_j \overline{N}_j.
\end{aligned}
$$

For events associated with type 1 individuals,

$$l_{(1,b)} = (0,1)^\top; l_{(1,d)} = (0,-1)^\top \text{ and}$$
$$\beta_{(1,b)}(\overline{\mathbf{N}}) = \lambda_1 \overline{N}_1; \beta_{(1,d)}(\overline{\mathbf{N}}) = \frac{\alpha \overline{N}_1}{\sum_{j=0}^1 \overline{N}_j} \sum_{j=0}^1 \lambda_j \overline{N}_j.$$

Since we only have finitely many events and $\beta_{(j,a)}$'s are continuous, we have for every compact $K \subset (0,\infty)^2$,

$$\sum_{(j,a)} ||l_{(j,a)}||_2 \sup_{\mathbf{x}\in K} \beta_{(j,a)}(\mathbf{x}) < \infty.$$

Since $\beta_{(j,b)}$, $\beta_{(j,m)}$, and $\beta_{(j,d)}$ are continuously differentiable on $(0,\infty)^2$, $\mathbf{F}$ is locally Lipschitz. Hence, by Theorem 2.1 in Ethier and Kurtz [2009], we have as $r \to \infty$,

$$\overline{\mathbf{N}}^{(r)} \to \overline{\mathbf{N}} \text{ almost surely.}$$

In the proof, we omit one state $(0,0)^\top$ as it has probability tending to zero as $r \to \infty$. □

Proposition 2. *As $r \to \infty$, $\widehat{\mathbf{N}}^{(r)} \Rightarrow \widehat{\mathbf{N}}$, where $\widehat{\mathbf{N}}$ satisfies*

$$d\widehat{\mathbf{N}}(t) = \nabla \mathbf{F}(\overline{\mathbf{N}}(t))\widehat{\mathbf{N}}(t)dt + \sigma(t)d\mathbf{B}(t); \qquad \widehat{\mathbf{N}}(0) = 0.$$

$\mathbf{B}$ is a vector of standard Brownian motions and $\sigma(t)$ is a 2 by 5 matrix. The expression for $\sigma(t)$ is

$$\sigma^\top(t) = \begin{bmatrix} \sqrt{(w_0+v_0)\overline{N}_0} & 0 \\ -\sqrt{v_0\overline{N}_0} & \sqrt{v_0\overline{N}_0} \\ -\sqrt{\frac{\alpha\overline{N}_0}{\sum_{j=0}^1 \overline{N}_j}\sum_{j=0}^1 (w_j+v_j)\overline{N}_j} & 0 \\ 0 & \sqrt{(w_1+v_1)\overline{N}_1} \\ 0 & -\sqrt{\frac{\alpha\overline{N}_1}{\sum_{j=0}^1 \overline{N}_j}\sum_{j=0}^1 (w_j+v_j)\overline{N}_j} \end{bmatrix}.$$

Proof. We verify conditions in Theorem 2.3 on page 458 of Ethier and Kurtz [2009]. Recall $l_{(j,a)}$'s and $\beta_{(j,a)}$'s defined in Proposition 1. By continuity of $\beta_{(j,a)}$'s, we have for every compact $K \subset (0,\infty)^2$,

$$\sum_{(j,a)} ||l_{(j,a)}||_2^2 \sup_{\mathbf{x}\in K} \beta_{(j,a)}(\mathbf{x}) < \infty.$$

Since $\beta_{(j,a)}$ is continuously differentiable for all (j,a)'s, $\nabla\mathbf{F}$ is continuous. Therefore, $\widehat{\mathbf{N}}^{(r)} \Rightarrow \widehat{\mathbf{N}}$ as $r \to \infty$, where $\widehat{\mathbf{N}}$ is defined by the stochastic differential equation in the statement of this Proposition. □

C Ratio Dynamics

For $t \geq 0$, we define the ratio dynamics as

$$R_1(t) := \frac{\overline{N}_1(t)}{\overline{N}_0(t)}; \quad t \geq 0.$$

Lemma 1. *For $t \geq 0$, we have*

$$R_1(t) = \begin{cases} v_0 t, & \text{if } w_0 = w_1 \\ \frac{v_0}{w_1 - w_0}(e^{(w_1-w_0)t} - 1), & \text{if } w_0 \neq w_1. \end{cases}$$

Proof. From 1, we have

$$\overline{N}_0'(t) = w_0\overline{N}_0(t) - \alpha\frac{\overline{N}_0(t)}{\sum_{j=0}^{1}\overline{N}_j(t)}\sum_{j=0}^{1}\lambda_j\overline{N}_j(t);$$

$$\overline{N}_1'(t) = w_1\overline{N}_1(t) + v_0\overline{N}_0(t) - \alpha\frac{\overline{N}_1(t)}{\sum_{j=0}^{1}\overline{N}_j(t)}\sum_{j=0}^{1}\lambda_j\overline{N}_j(t),$$

with initial condition $\overline{\mathbf{N}}(0) = (1,0)^\top$.

Therefore,

$$R_1' = \frac{\overline{N}_1'\overline{N}_0 - \overline{N}_1\overline{N}_0'}{\overline{N}_0^2} = \frac{w_1\overline{N}_0\overline{N}_1 + v_0\overline{N}_0^2 - w_0\overline{N}_0\overline{N}_1}{\overline{N}_0^2} = (w_1 - w_0)R_1 + v_0.$$

This completes the proof. □

D Approximated Dynamics of VAF

For $t \geq 0$, we define the VAF dynamics as

$$\overline{P}^{(r)}(t) = \frac{1}{2}\frac{\overline{N}_1^{(r)}(t)}{\overline{N}_0^{(r)}(t) + \overline{N}_1^{(r)}(t)}.$$

Theorem 1. $\overline{P}_1^{(r)} \to \overline{P}_1$ *almost surely, where*

$$\overline{P}_1(t) = \begin{cases} \frac{1}{2}\cdot\frac{v_0 t}{1+v_0 t}, & \text{if } w_0 = w_1; \\ \frac{1}{2}\cdot\frac{v_0[e^{(w_1-w_0)t}-1]}{w_1-w_0+v_0[e^{(w_1-w_0)t}-1]}, & \text{if } w_0 < w_1. \end{cases}$$

Proof. We define $g : \mathcal{D}^2 \to \mathcal{D}$ such that for all $x, y \in \mathcal{D}$,

$$g(x,y) = \frac{1}{2}\frac{y}{x+y}.$$

It is evident that $g(x,y)$ is an element of $\mathcal{D}$. Usually, the addition operator is not continuous on space $\mathcal{D}$. That is, there exists $x_n \to x$ and $y_n \to y$ in $\mathcal{D}$, but $x_n + y_n \nrightarrow x + y$. For a concrete example, see Example 3.3.1 in Whitt [2002]. Nonetheless, if both x and y are continuous, we do have $\lim_{n\to\infty}(x_n + y_n) = x + y$. Using similar reasoning, if both x and y are continuous, the ratio operator is also continuous. Hence, $g(x_n, y_n)$ converges to $g(x,y)$ in $\|\cdot\|_T$ for all $T > 0$. This concludes that g is continuous at $\overline{\mathbf{N}}$ and by the continuous mapping theorem, we have

$$g(\overline{\mathbf{N}}^{(r)}) \to g(\overline{\mathbf{N}}) \text{ almost surely.}$$

□

Theorem 2. *As $r \to \infty$, the finite dimensional distributions of $\widehat{P}_1^{(r)}$ converges weakly to the finite dimensional distribution of $\widehat{P}_1$ defined by*

$$\widehat{P}_1(t) = \frac{1}{[\overline{N}_0(t) + \overline{N}_1(t)]}\begin{bmatrix} -\overline{P}_1(t) & \overline{P}_0(t) \end{bmatrix}\widehat{\mathbf{N}}(t).$$

Proof. This is demonstrated by the delta method. Since as $r \to \infty$,

$$\sqrt{r}(\overline{\mathbf{N}}^{(r)} - \overline{\mathbf{N}}) \Rightarrow \widehat{\mathbf{N}} \sim GM(\mathbf{0}, \rho(\cdot,\cdot)),$$

we can derive its finite dimensional distribution on $(t_1, \cdots, t_n)$, which is a multivariate Gaussian distribution with mean being the zero vector and the covariance matrix S being

$$S = \begin{bmatrix} \rho(t_1,t_1) & \rho(t_1,t_2) & \cdots & \rho(t_1,t_n) \\ \rho(t_2,t_1) & \rho(t_2,t_2) & \cdots & \rho(t_2,t_n) \\ \vdots & \vdots & \ddots & \vdots \\ \rho(t_n,t_1) & \rho(t_n,t_2) & \cdots & \rho(t_n,t_n) \end{bmatrix}.$$

Define a function $h : \mathbb{R}^{2n} \to \mathbb{R}^n$ such that

$$h(x_1, y_1, x_2, y_2, \cdots, x_n, y_n) = (\frac{y_1}{x_1 + y_1}, \cdots, \frac{y_n}{x_n + y_n})^\top. \tag{23}$$

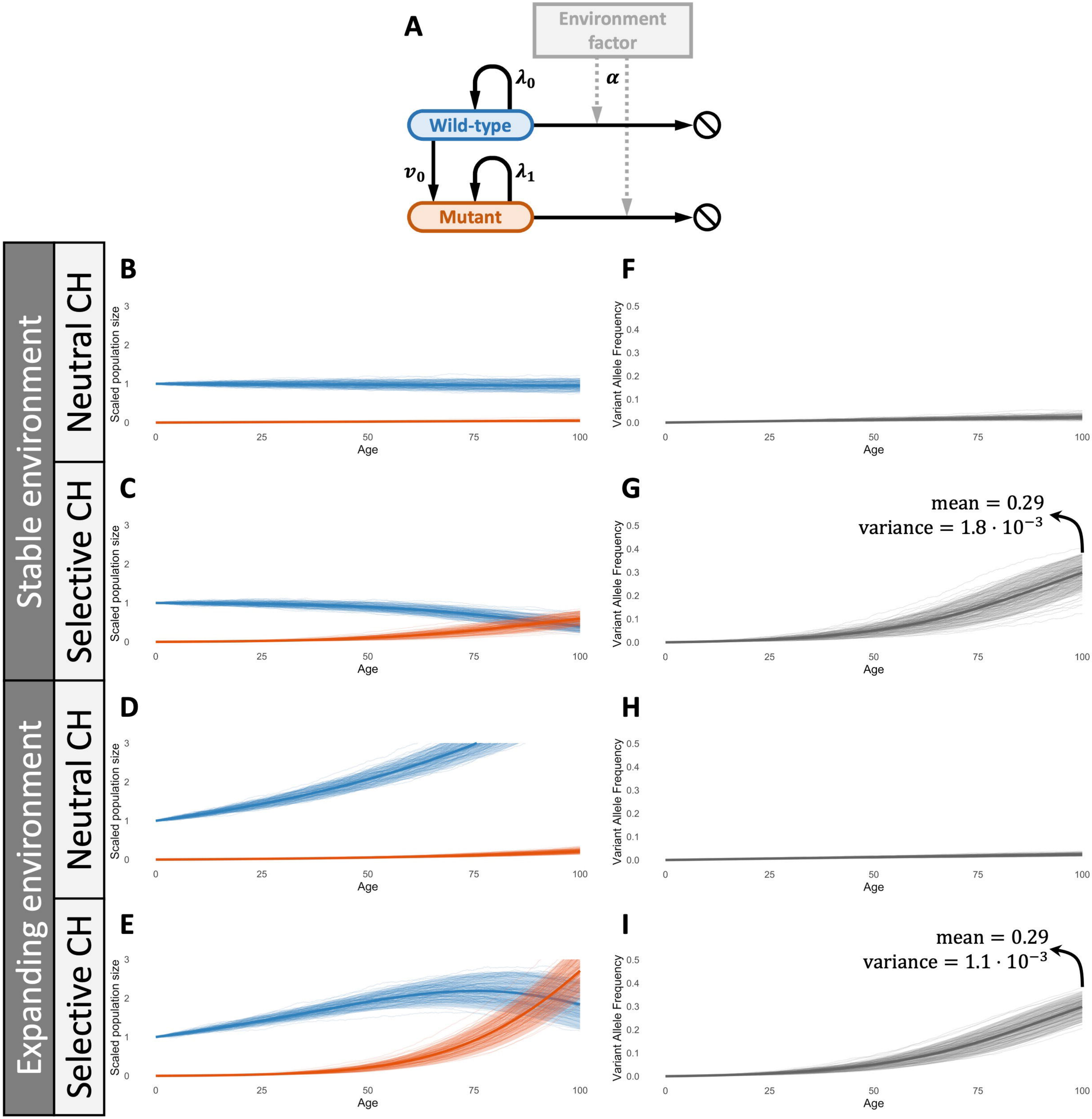
Therefore, the Jacobian of h is

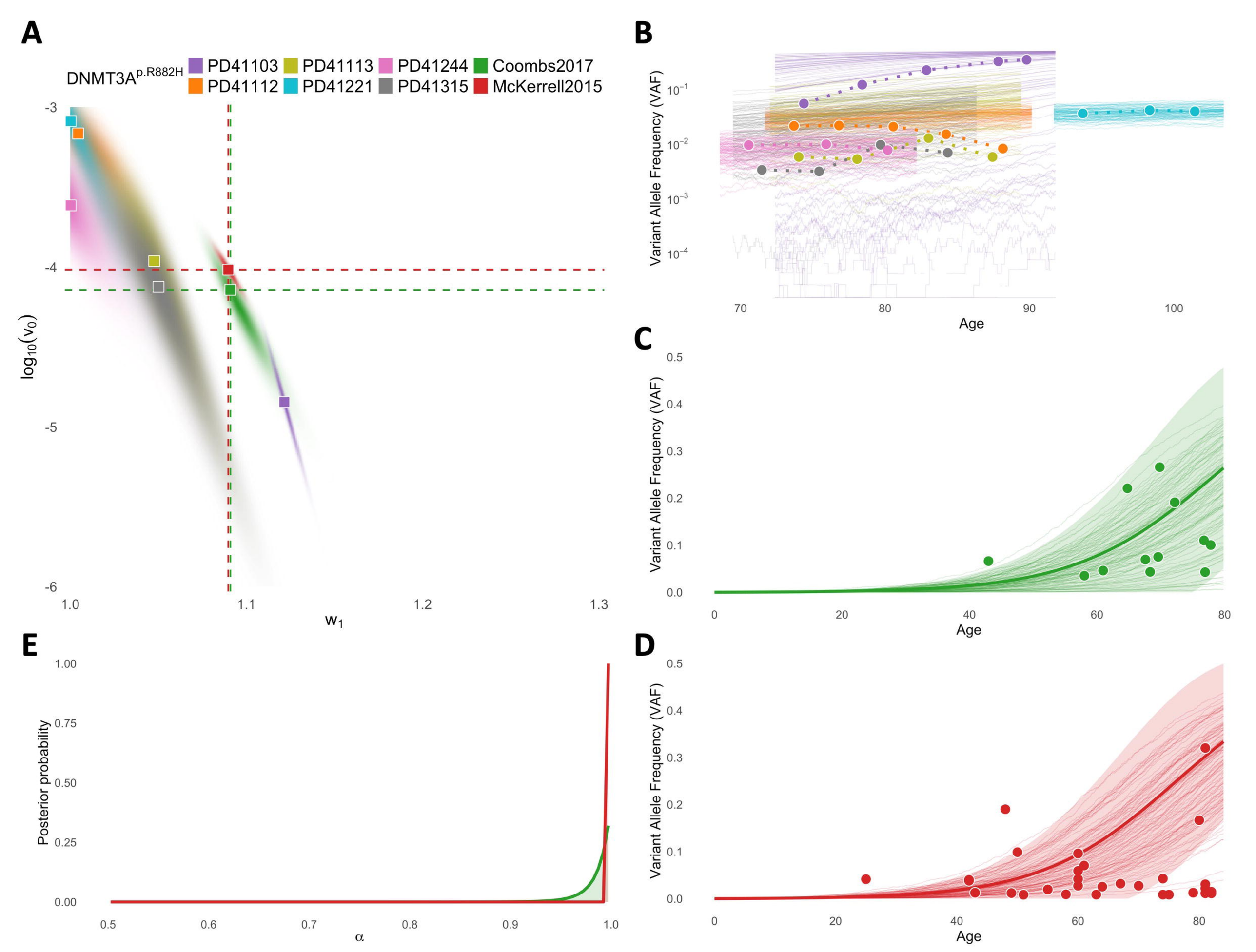
$$\nabla h(x_1, y_1, x_2, y_2, \cdots, x_n, y_n) = \begin{bmatrix} \frac{-y_1}{(x_1+y_1)^2} & \frac{x_1}{(x_1+y_1)^2} & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \frac{-y_2}{(x_2+y_2)^2} & \frac{x_2}{(x_2+y_2)^2} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \frac{-y_n}{(x_n+y_n)^2} & \frac{x_n}{(x_n+y_n)^2} \end{bmatrix}.$$

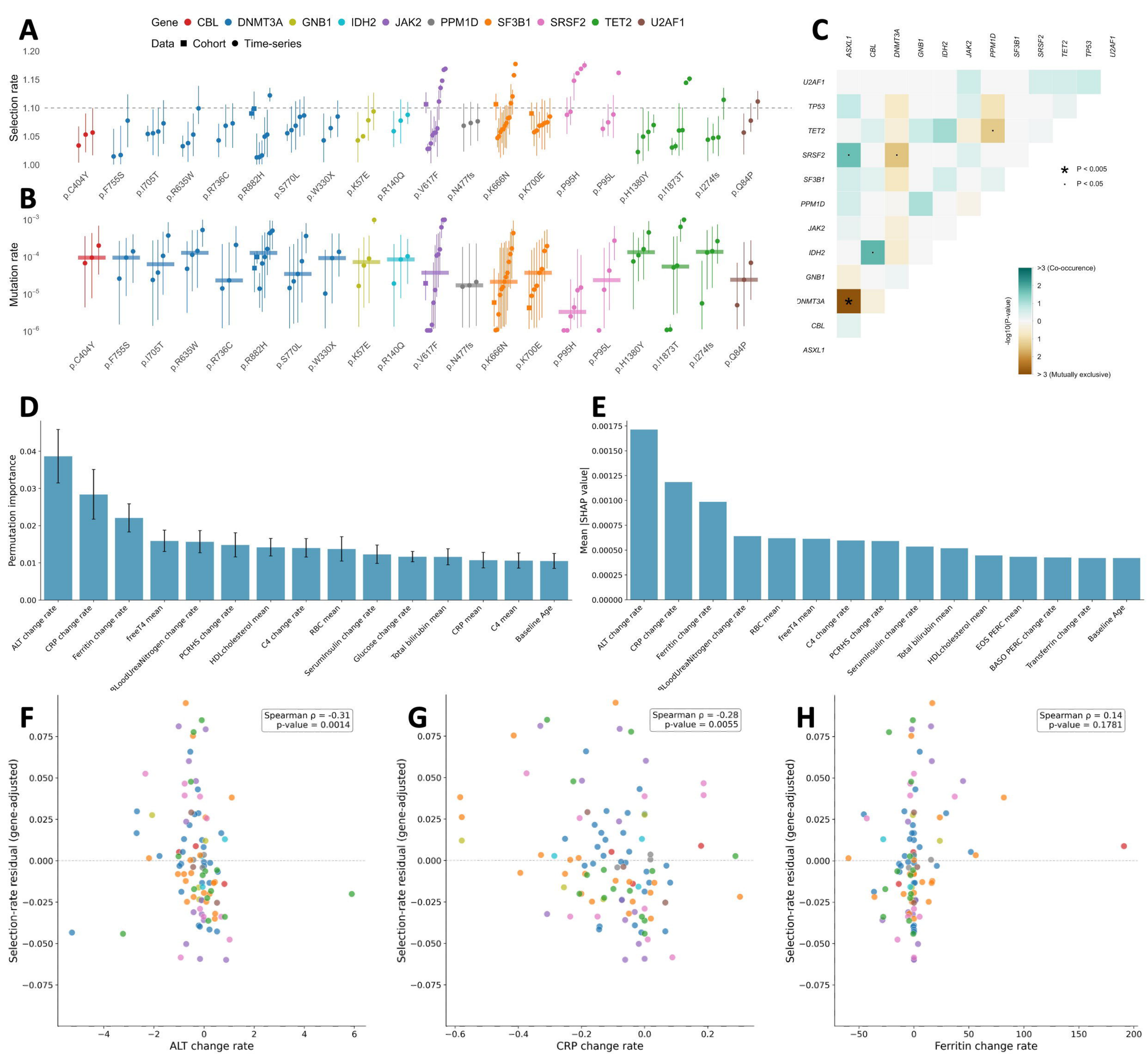
By delta method, we have

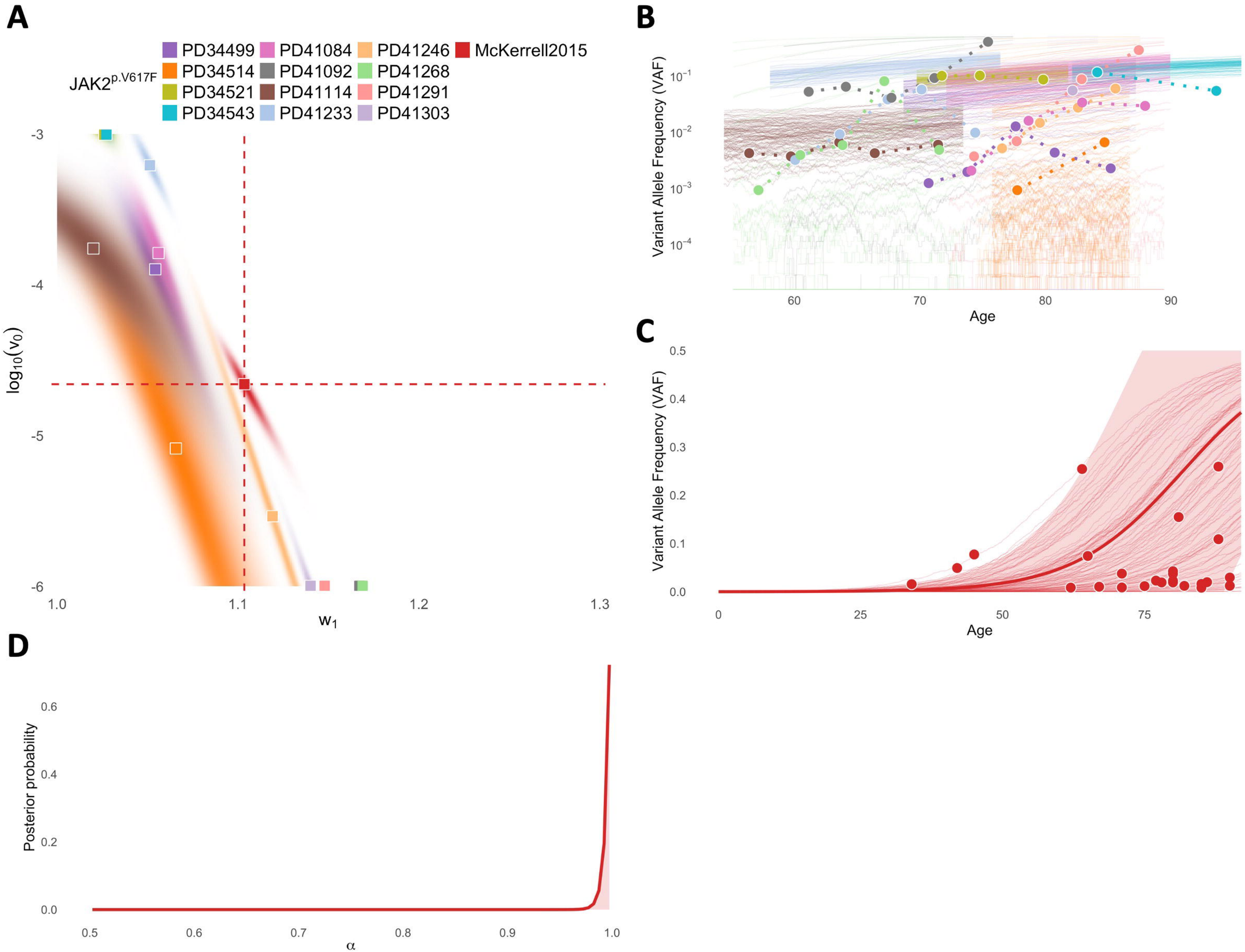
$$\sqrt{r}\left[\begin{bmatrix} \overline{P}_1^{(r)}(t_1) \\ \vdots \\ \overline{P}_1^{(r)}(t_n) \end{bmatrix} - \begin{bmatrix} \overline{P}_1(t_1) \\ \vdots \\ \overline{P}_1(t_n) \end{bmatrix}\right] \Rightarrow \nabla h(\overline{N}_0(t_1), \overline{N}_1(t_1), \cdots, \overline{N}_0(t_n), \overline{N}_1(t_n)) \begin{bmatrix} \widehat{N}_0(t_1) \\ \widehat{N}_1(t_1) \\ \vdots \\ \widehat{N}_0(t_n) \\ \widehat{N}_1(t_n) \end{bmatrix}.$$

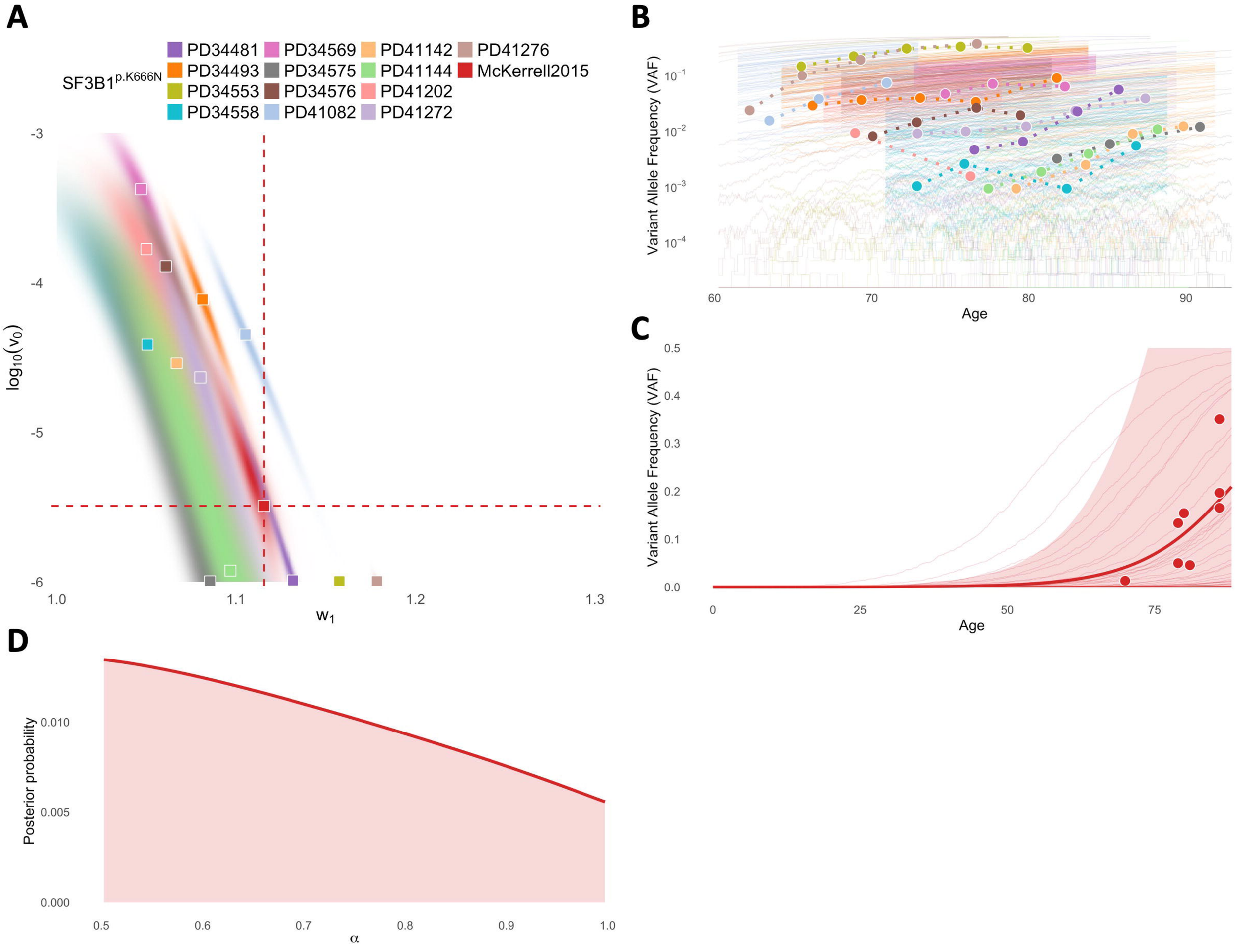
Hence, finite dimensional distribution of $\widehat{P}_1^{(r)}$ converges to that of $\widehat{P}_1$. □

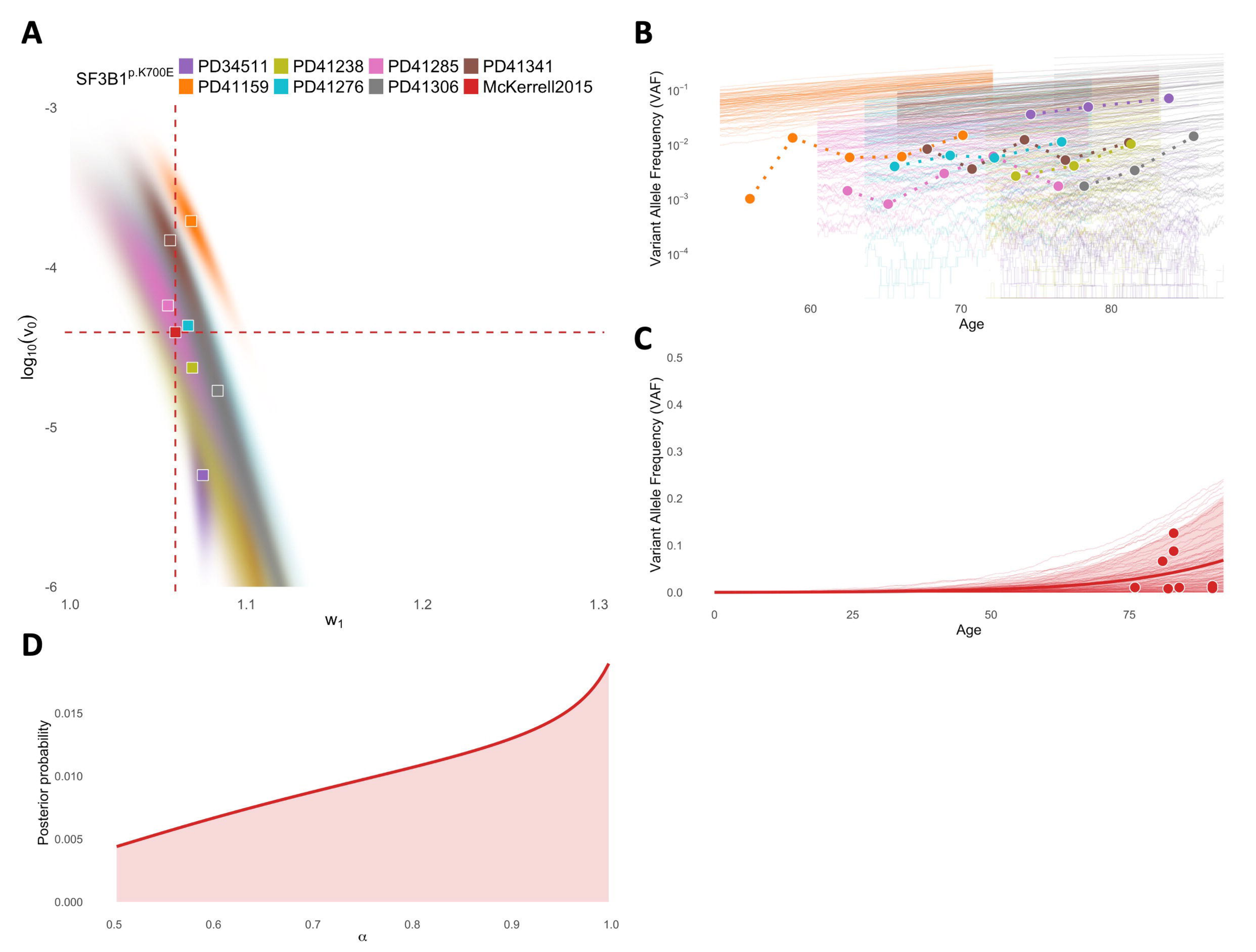

A
Environment factor
λ_0
α
Wild-type
v_0
λ_1
Mutant
Stable environment
Expanding environment
Neutral CH
Selective CH
Neutral CH
Selective CH
B
C
D
E
F
G
H
I
Scaled population size
Variant Allele Frequency
Age
mean = 0.29
variance = $1.8 \cdot 10^{-3}$
mean = 0.29
variance = $1.1 \cdot 10^{-3}$

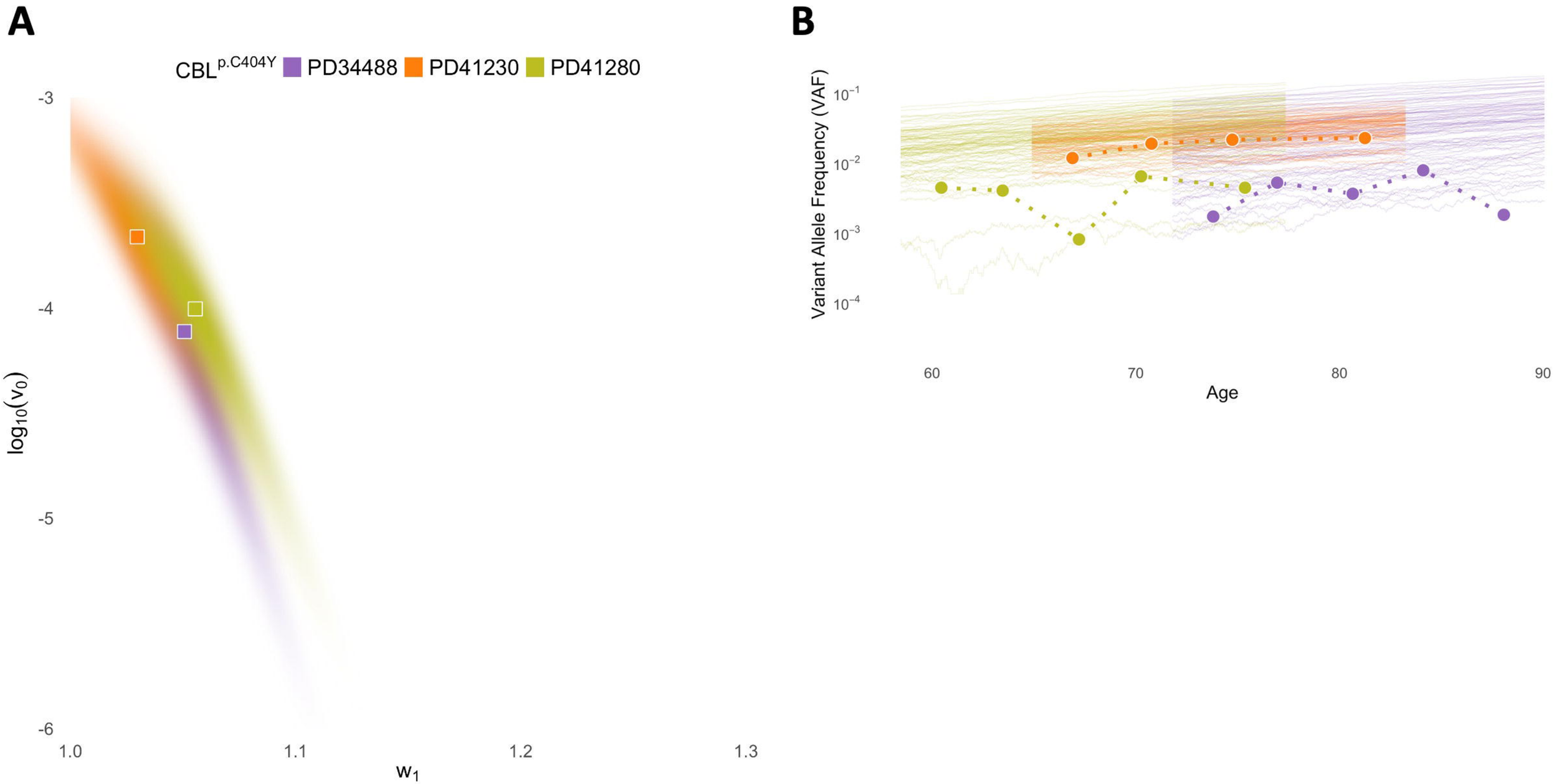

A
DNMT3A p.R882H
PD41103
PD41112
PD41113
PD41221
PD41244
PD41315
Coombs2017
McKerrell2015
log10(v0)
w1
B
Variant Allele Frequency (VAF)
Age
C
Variant Allele Frequency (VAF)
Age
D
Variant Allele Frequency (VAF)
Age
E
Posterior probability
α

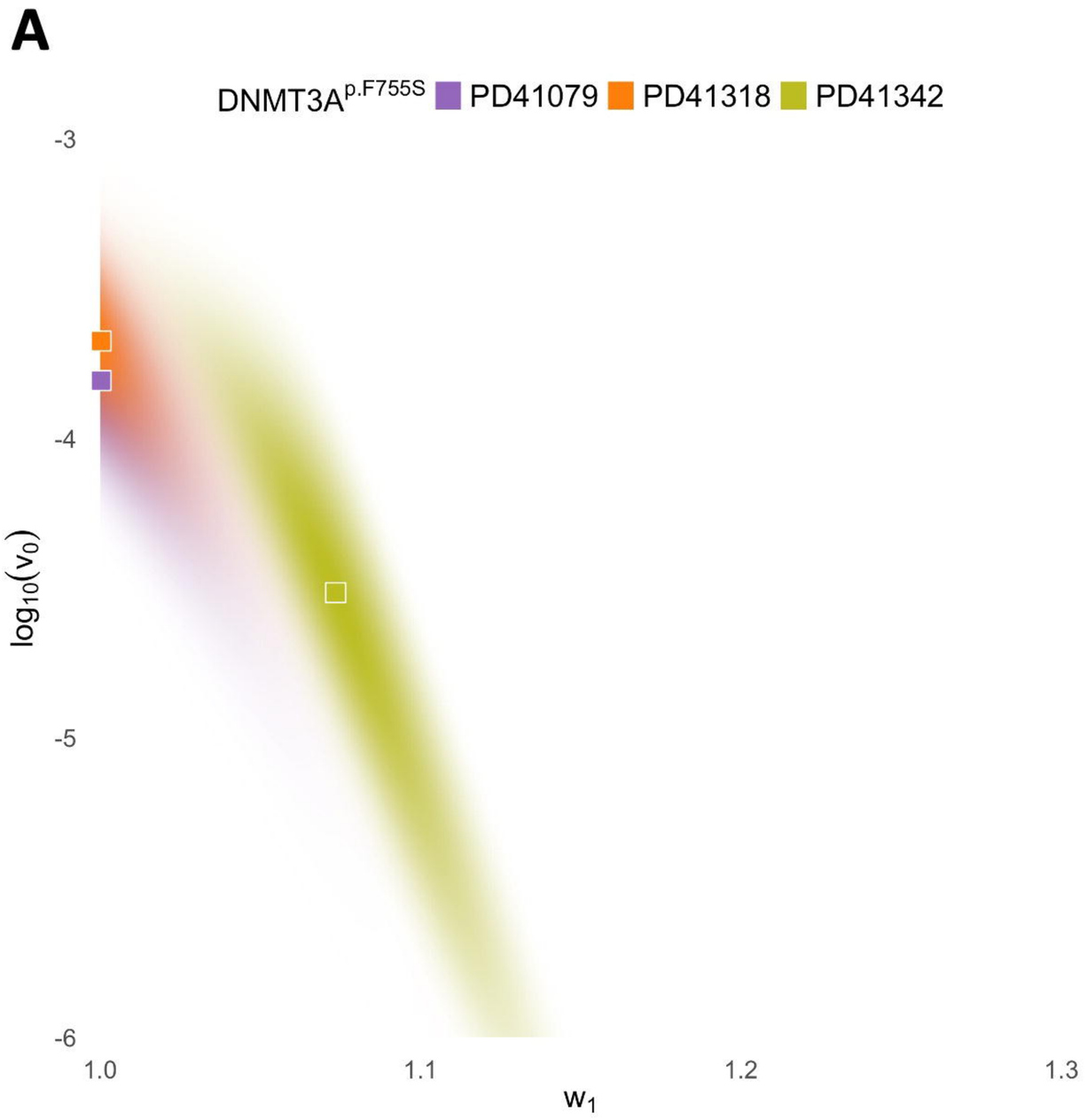

A
Gene CBL DNMT3A GNB1 IDH2 JAK2 PPM1D SF3B1 SRSF2 TET2 U2AF1
Data Cohort Time-series
Selection rate
p.C404Y p.F755S p.I705T p.R635W p.R736C p.R882H p.S770L p.W330X p.K57E p.R140Q p.V617F p.N477fs p.K666N p.K700E p.P95H p.P95L p.H1380Y p.I1873T p.I274fs p.Q84P
B
Mutation rate
C
ASXL1 CBL DNMT3A GNB1 IDH2 JAK2 PPM1D SF3B1 SRSF2 TET2 TP53 U2AF1
P < 0.005
P < 0.05
-log10(P-value)
>3 (Co-occurence)
> 3 (Mutually exclusive)
D
Permutation importance
ALT change rate
CRP change rate
Ferritin change rate
freeT4 mean
BloodUreaNitrogen change rate
PCRHS change rate
HDLcholesterol mean
C4 change rate
RBC mean
SerumInsulin change rate
Glucose change rate
Total bilirubin mean
CRP mean
C4 mean
Baseline Age
E
Mean |SHAP value|
ALT change rate
CRP change rate
Ferritin change rate
BloodUreaNitrogen change rate
RBC mean
freeT4 mean
C4 change rate
PCRHS change rate
SerumInsulin change rate
Total bilirubin mean
HDLcholesterol mean
EOS PERC mean
BASO PERC change rate
Transferrin change rate
Baseline Age
F
Spearman ρ = -0.31
p-value = 0.0014
Selection-rate residual (gene-adjusted)
ALT change rate
G
Spearman ρ = -0.28
p-value = 0.0055
Selection-rate residual (gene-adjusted)
CRP change rate
H
Spearman ρ = 0.14
p-value = 0.1781
Selection-rate residual (gene-adjusted)
Ferritin change rate

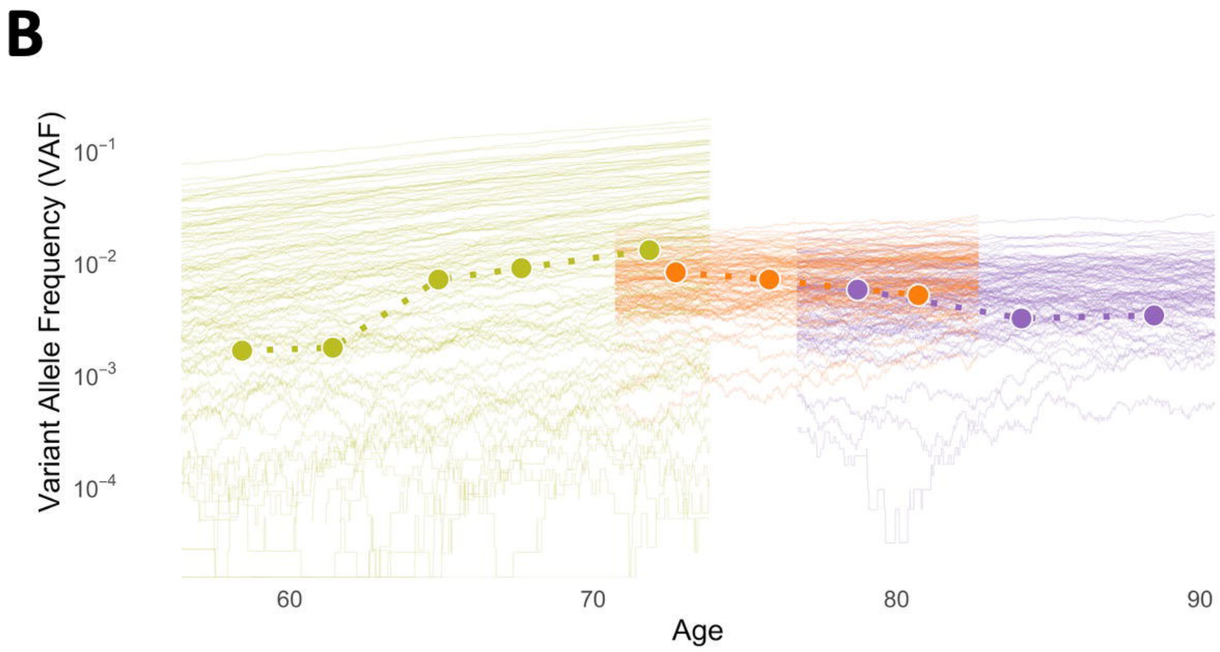

A
JAK2p.V617F
PD34499
PD34514
PD34521
PD34543
PD41084
PD41092
PD41114
PD41233
PD41246
PD41268
PD41291
PD41303
McKerrell2015
log10(v0)
w1
B
Variant Allele Frequency (VAF)
Age
C
Variant Allele Frequency (VAF)
Age
D
Posterior probability
α

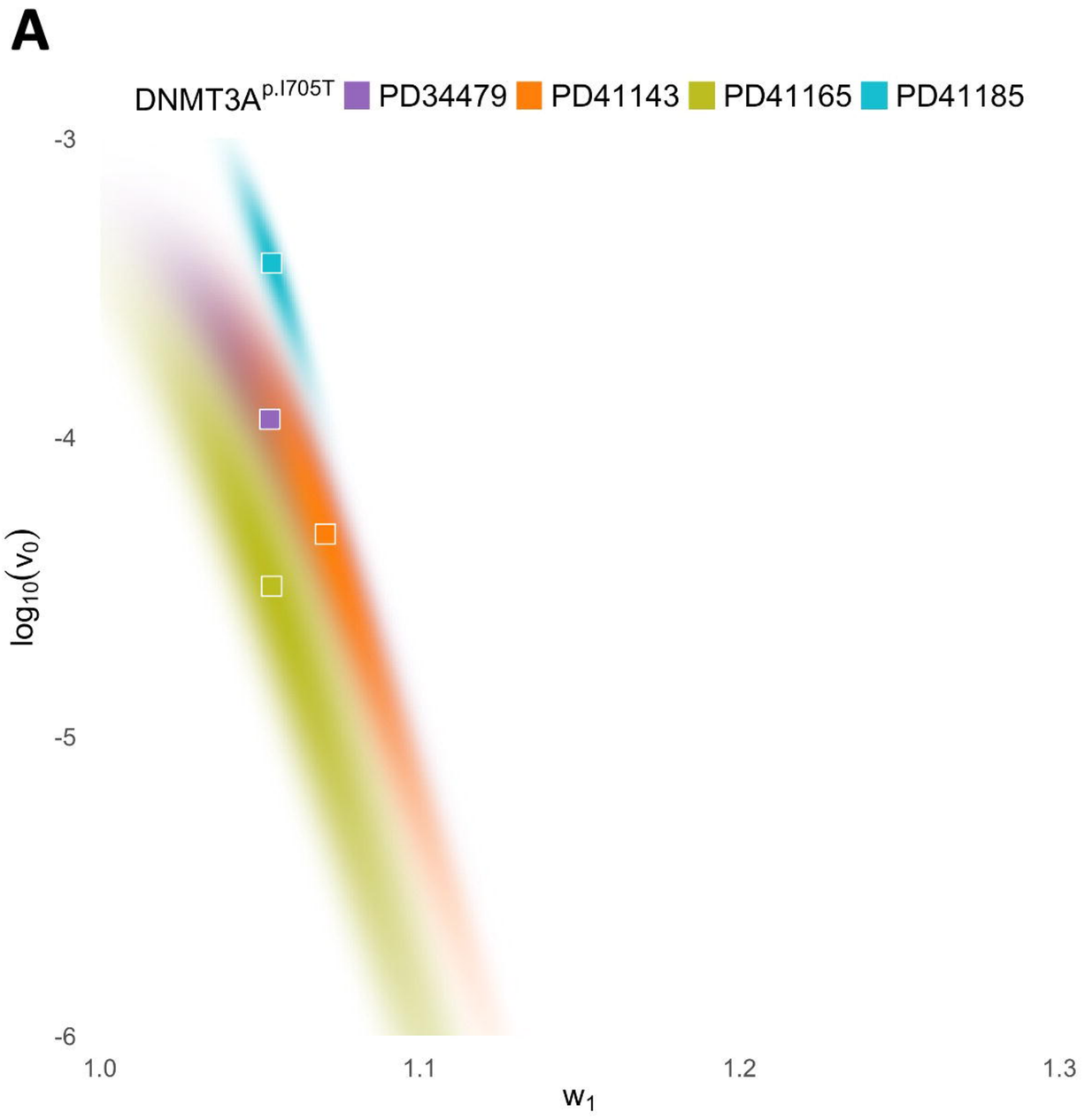

A
SF3B1p.K666N
PD34481
PD34493
PD34553
PD34558
PD34569
PD34575
PD34576
PD41082
PD41142
PD41144
PD41202
PD41272
PD41276
McKerrell2015
log10(v0)
w1
B
Variant Allele Frequency (VAF)
Age
C
Variant Allele Frequency (VAF)
Age
D
Posterior probability
α

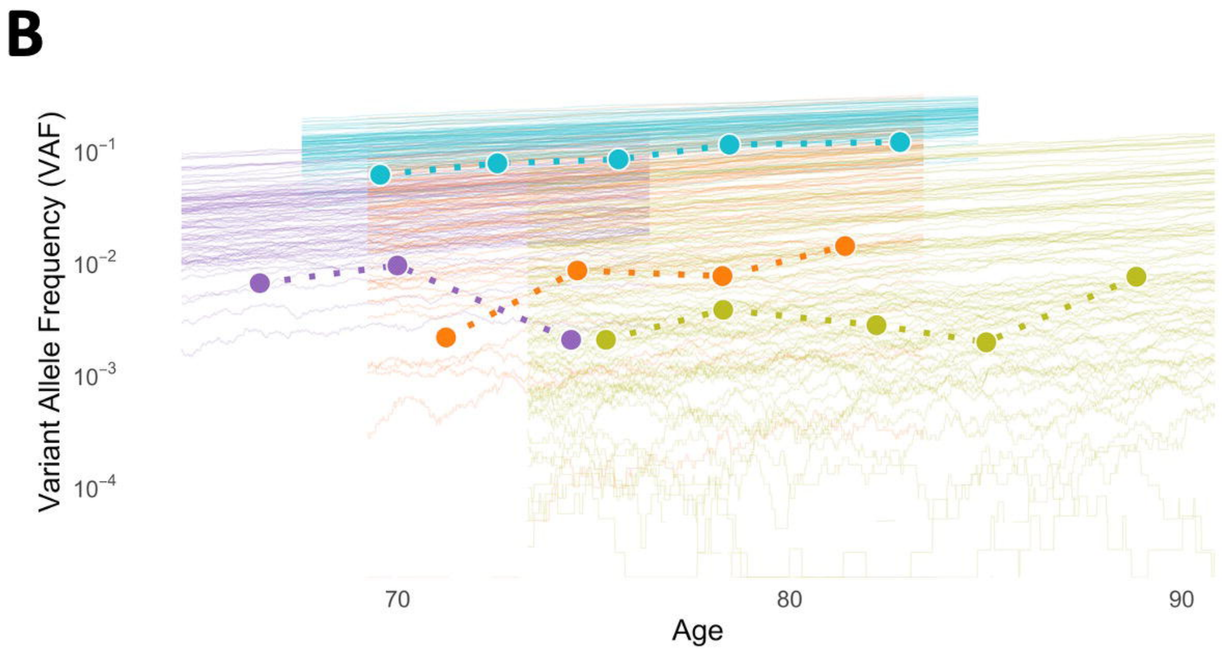

A
SF3B1p.K700E
PD34511
PD41238
PD41285
PD41341
PD41159
PD41276
PD41306
McKerrell2015
log10(v0)
w1
B
Variant Allele Frequency (VAF)
Age
C
Variant Allele Frequency (VAF)
Age
D
Posterior probability
α

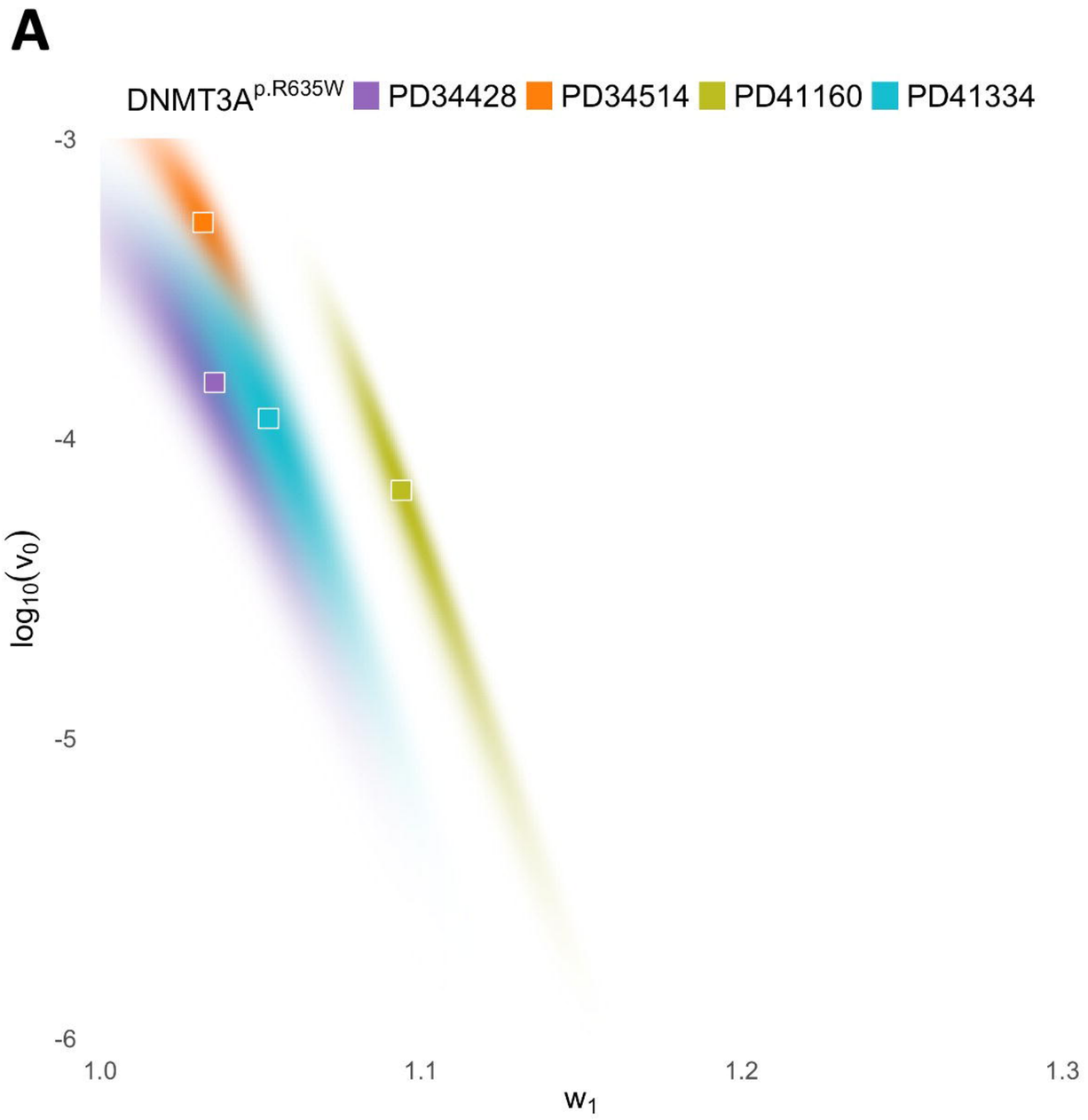

A
CBL p.C404Y
PD34488
PD41230
PD41280
-3
-4
-5
-6
log10(v0)
1.0
1.1
1.2
1.3
w1
B
Variant Allele Frequency (VAF)
10−1
10−2
10−3
10−4
60
70
80
90
Age

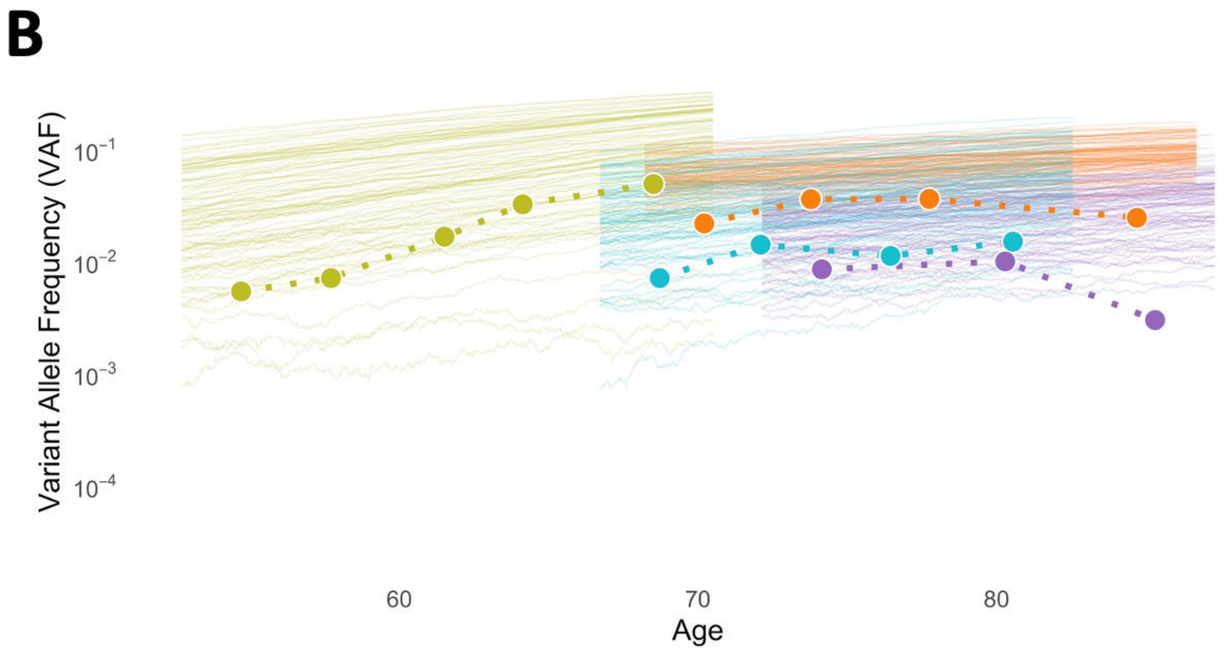

A
DNMT3A p.F755S
PD41079
PD41318
PD41342
-3
-4
-5
-6
log10(v0)
1.0
1.1
1.2
1.3
w1

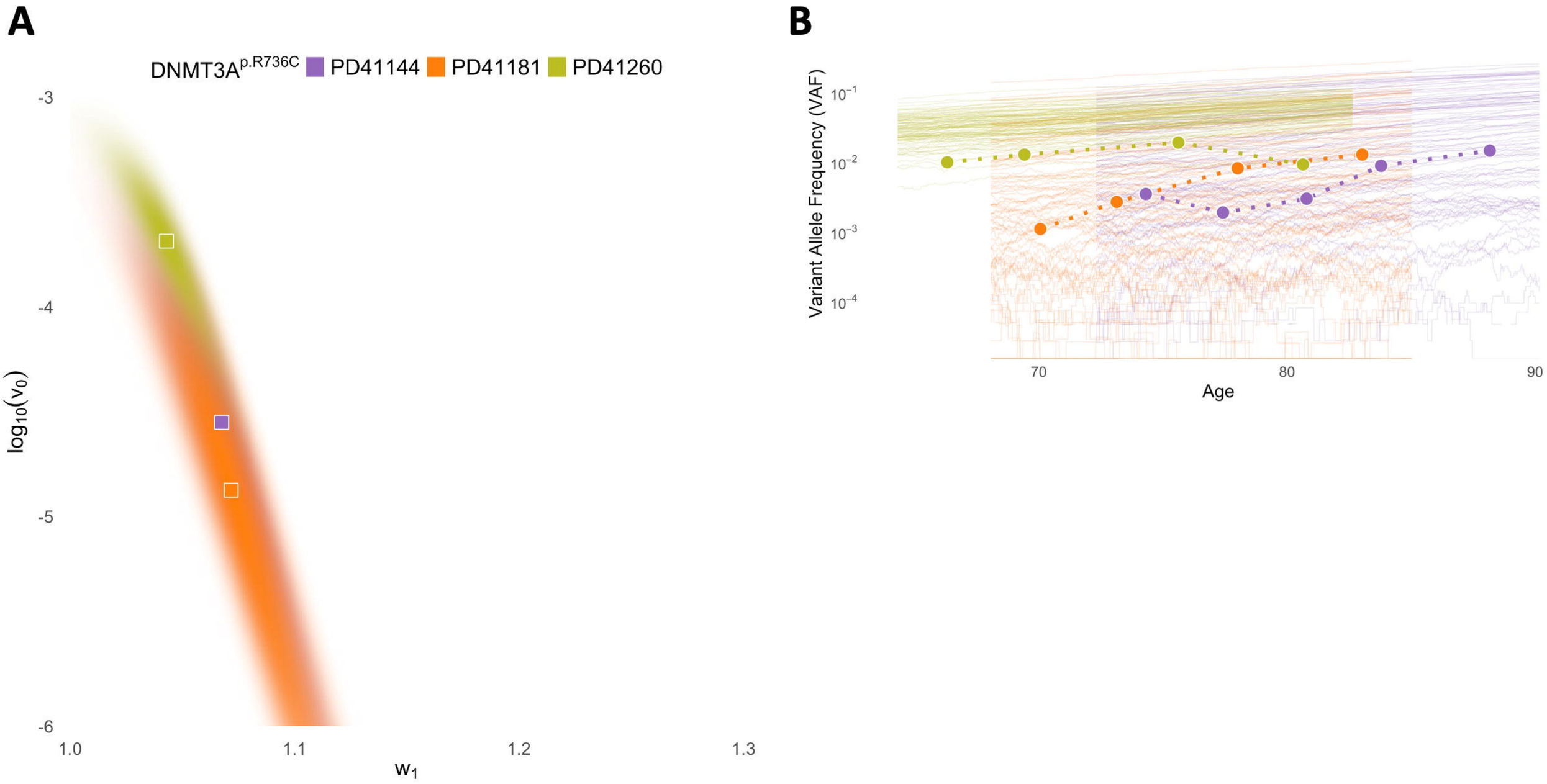

B
Variant Allele Frequency (VAF)
10−1
10−2
10−3
10−4
60
70
80
90
Age

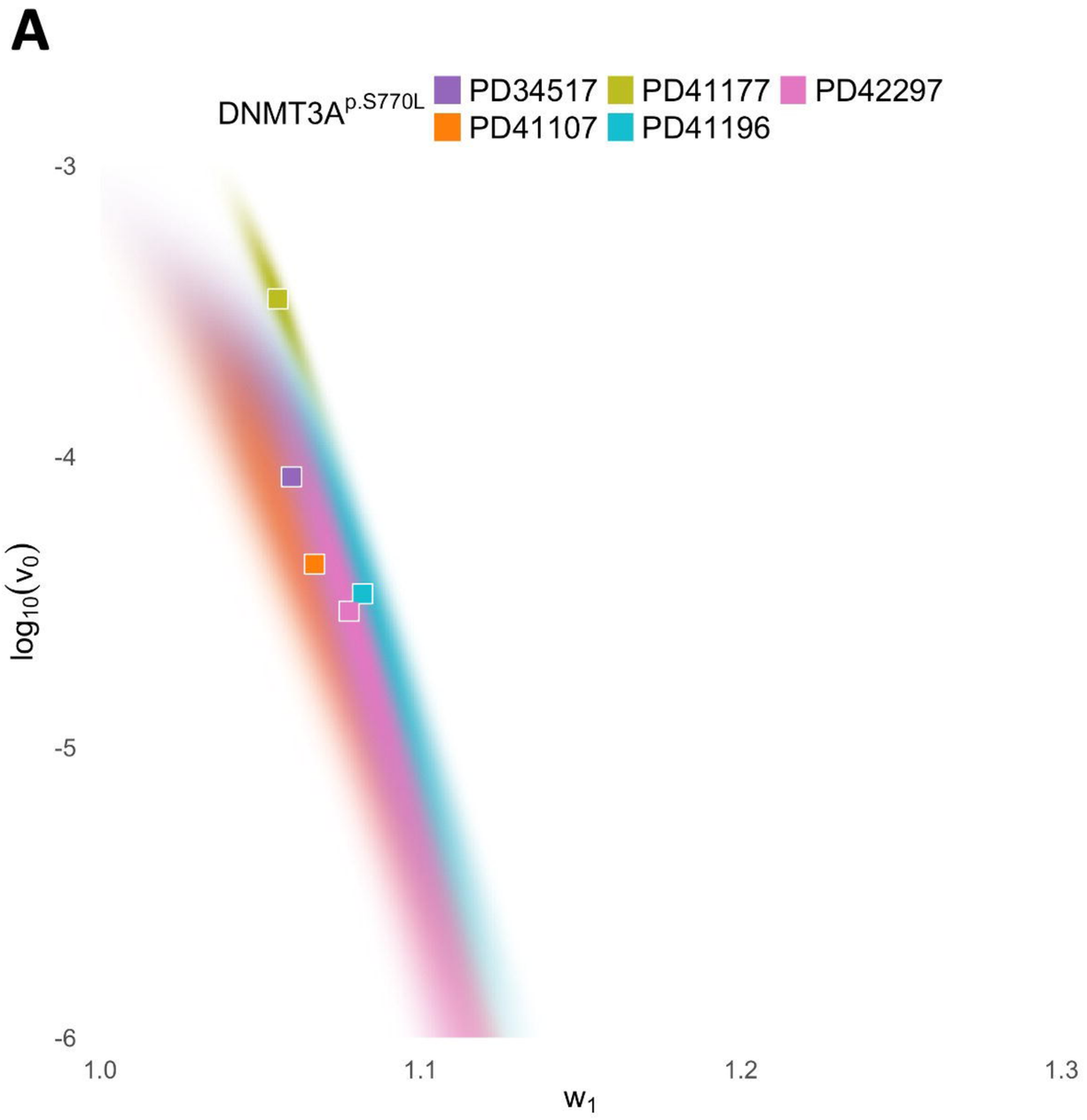

A
DNMT3A p.I705T
PD34479
PD41143
PD41165
PD41185
-3
-4
-5
-6
log10(v0)
1.0
1.1
1.2
1.3
w1

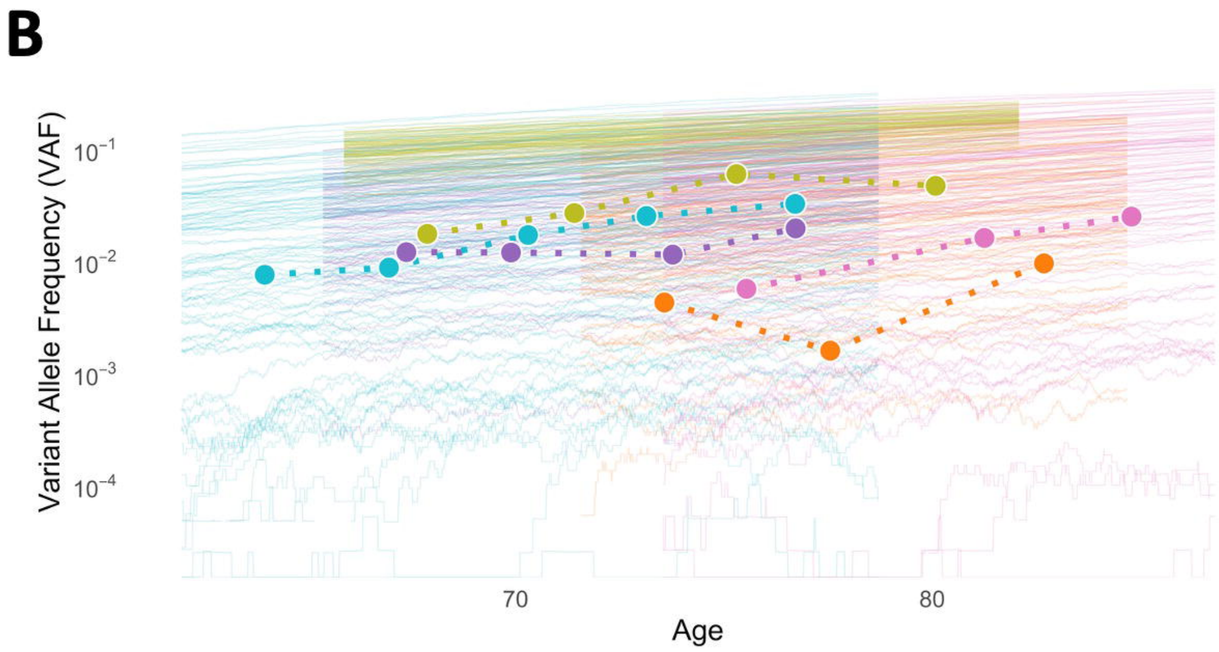

B
Variant Allele Frequency (VAF)
10−1
10−2
10−3
10−4
70
80
90
Age

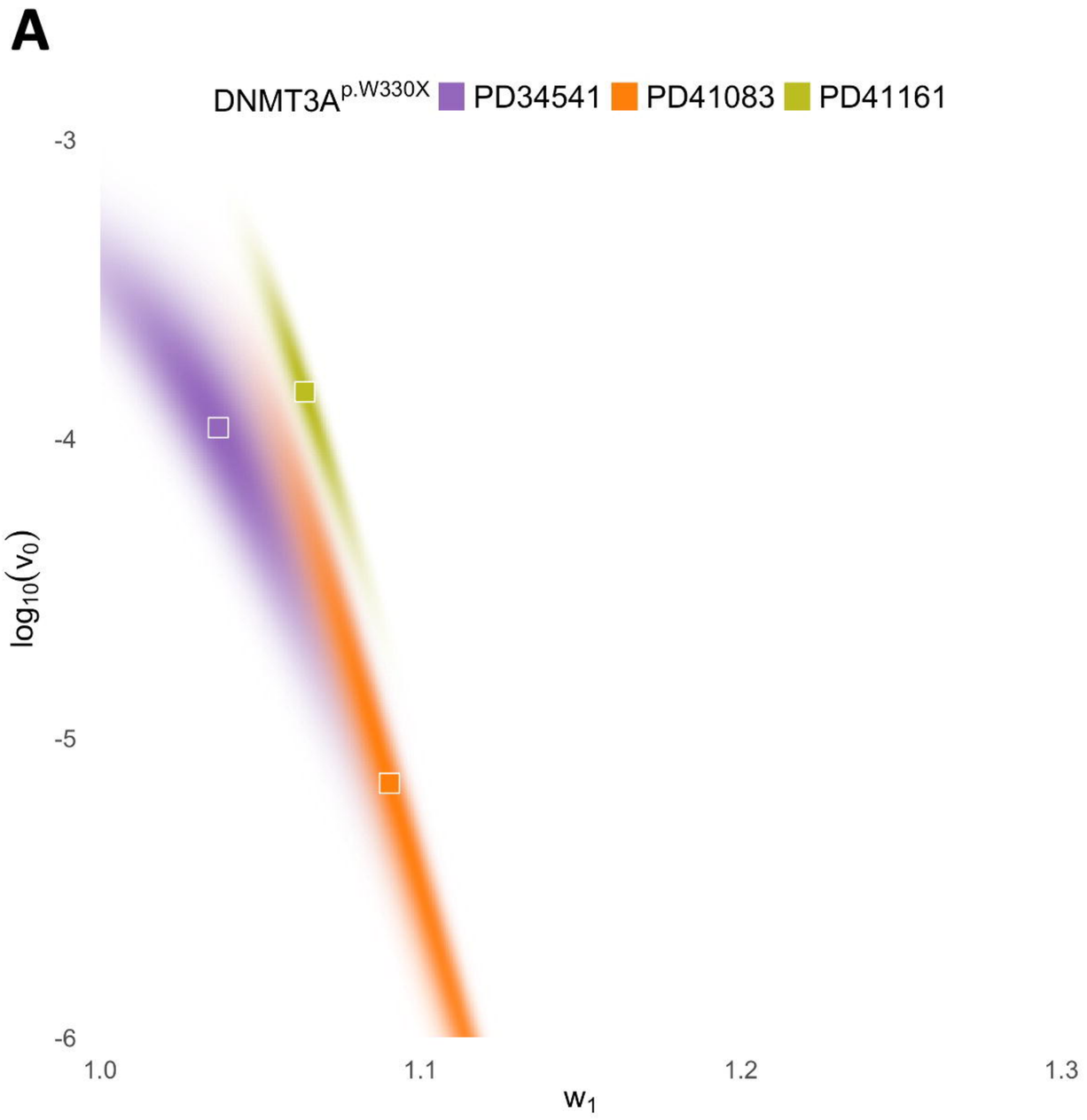

A
DNMT3A p.R635W
PD34428
PD34514
PD41160
PD41334
-3
-4
-5
-6
log10(v0)
1.0
1.1
1.2
1.3
w1

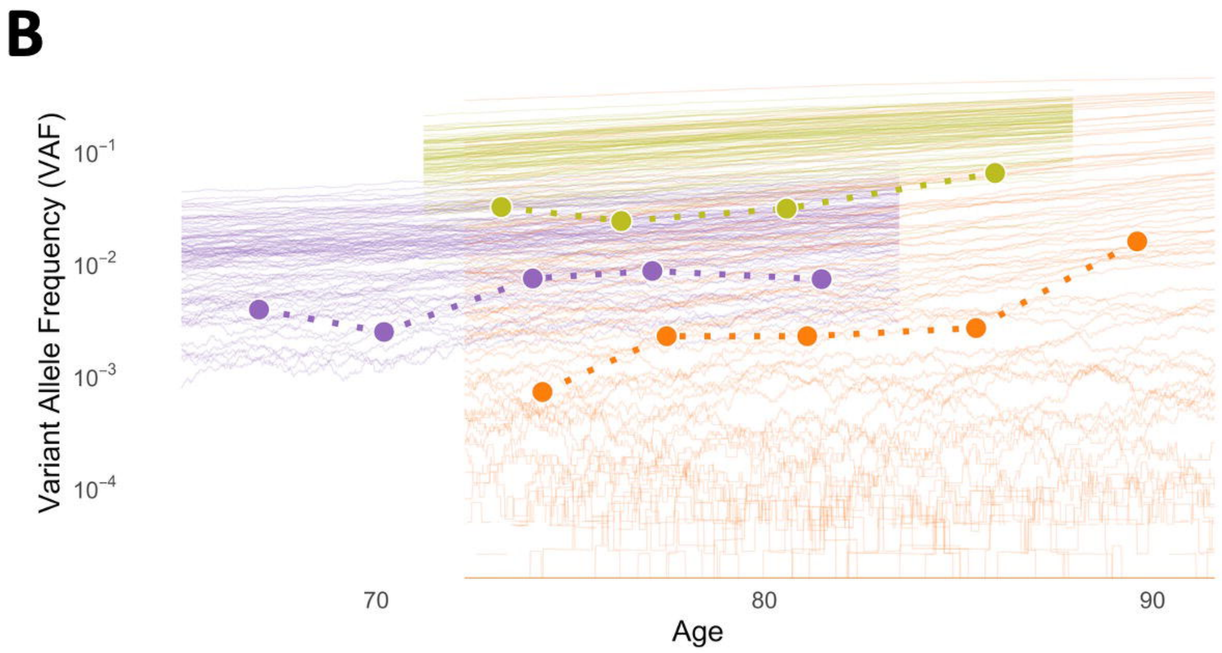

B
Variant Allele Frequency (VAF)
10−1
10−2
10−3
10−4
60
70
80
Age

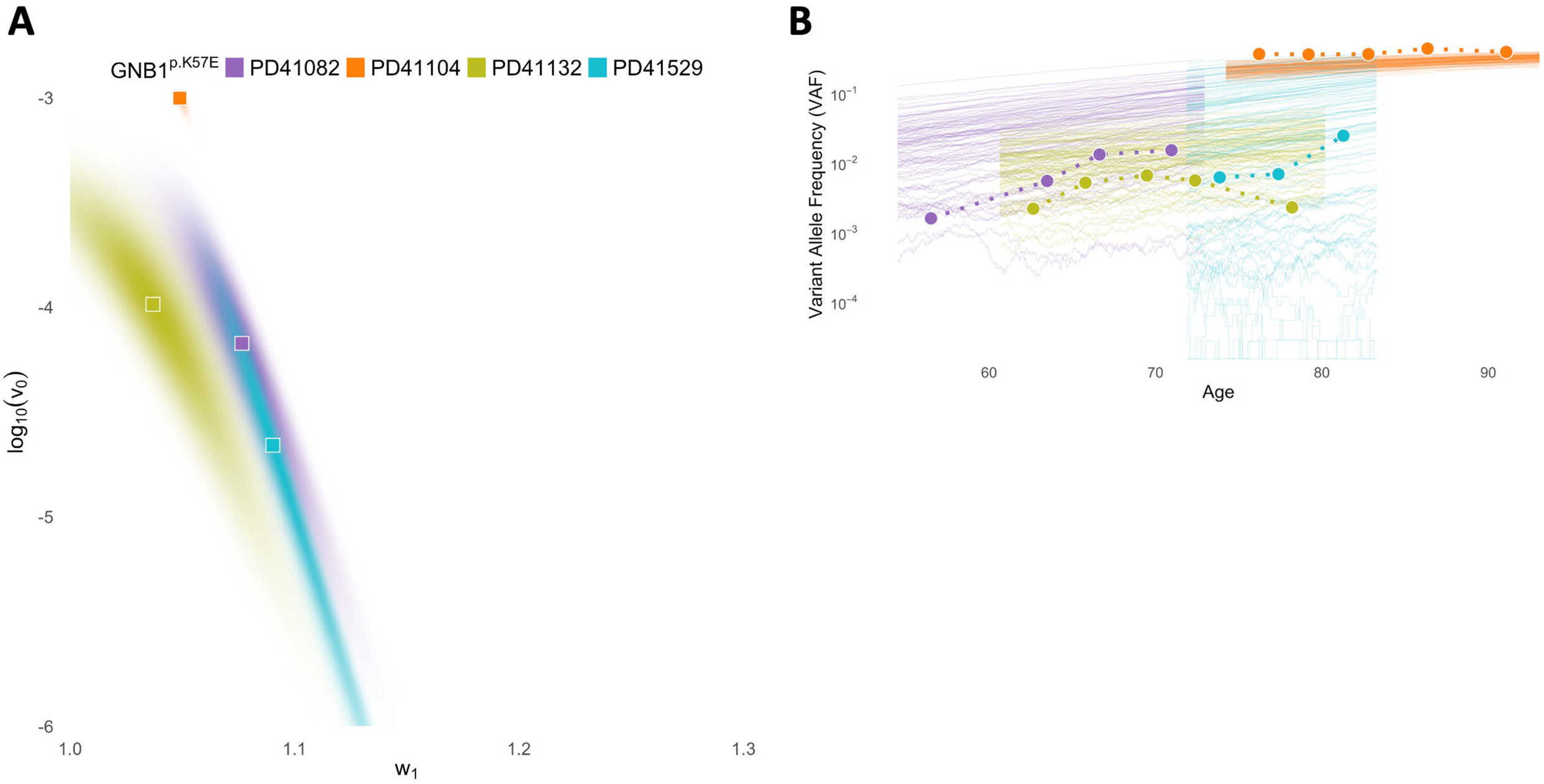

A
DNMT3A p.R736C
PD41144
PD41181
PD41260
-3
-4
-5
-6
log10(v0)
1.0
1.1
1.2
1.3
w1
B
Variant Allele Frequency (VAF)
10−1
10−2
10−3
10−4
70
80
90
Age

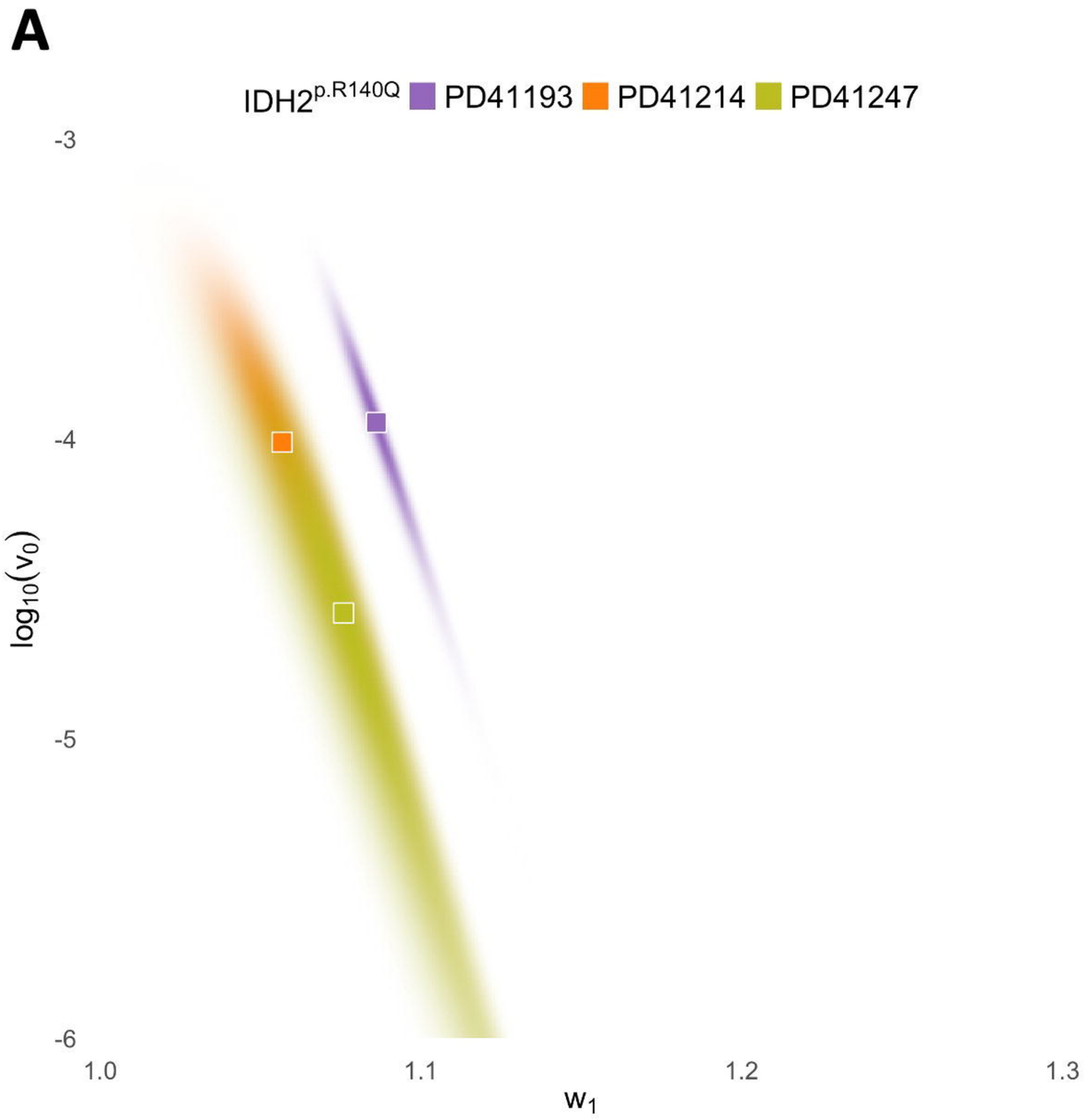

A
DNMT3A$^{p.S770L}$
PD34517
PD41177
PD42297
PD41107
PD41196
-3
-4
-5
-6
$\log_{10}(v_0)$
1.0
1.1
1.2
1.3
w_1

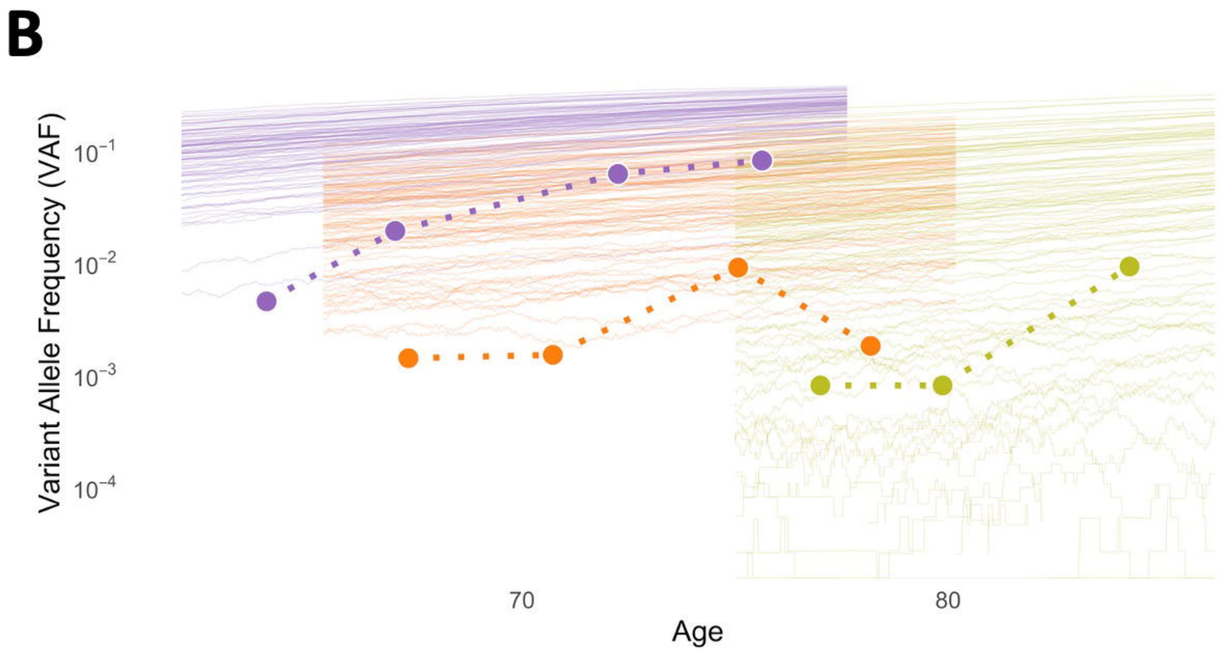

B
Variant Allele Frequency (VAF)
10^{-1}
10^{-2}
10^{-3}
10^{-4}
70
80
Age

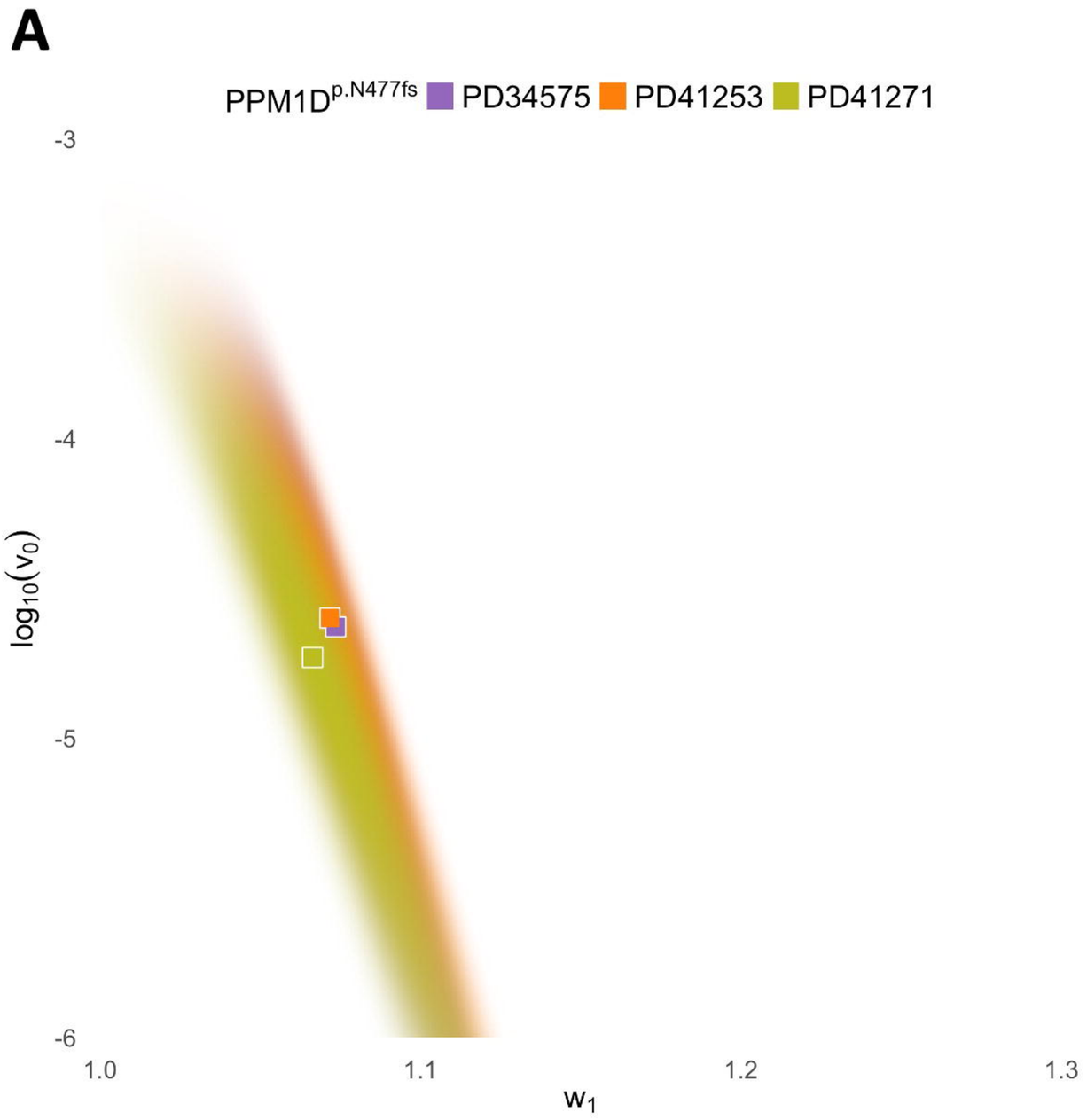

A
DNMT3A p.W330X
PD34541
PD41083
PD41161
-3
-4
-5
-6
log10(v0)
1.0
1.1
1.2
1.3
w1

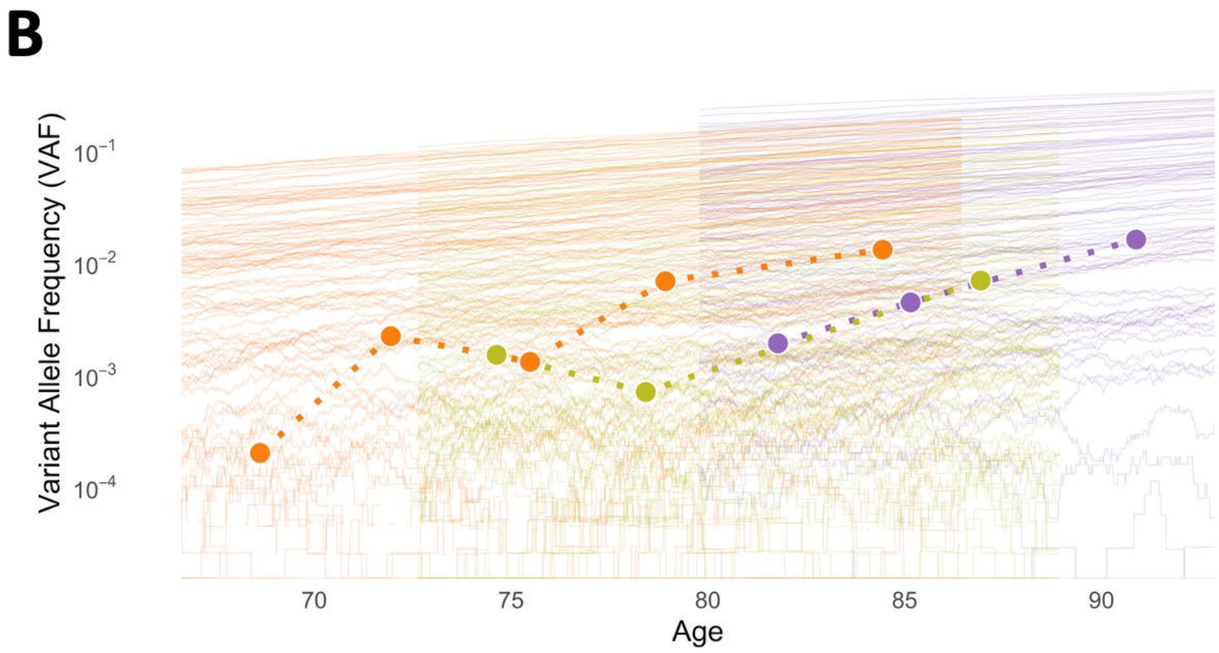

B
Variant Allele Frequency (VAF)
10−1
10−2
10−3
10−4
70
80
90
Age

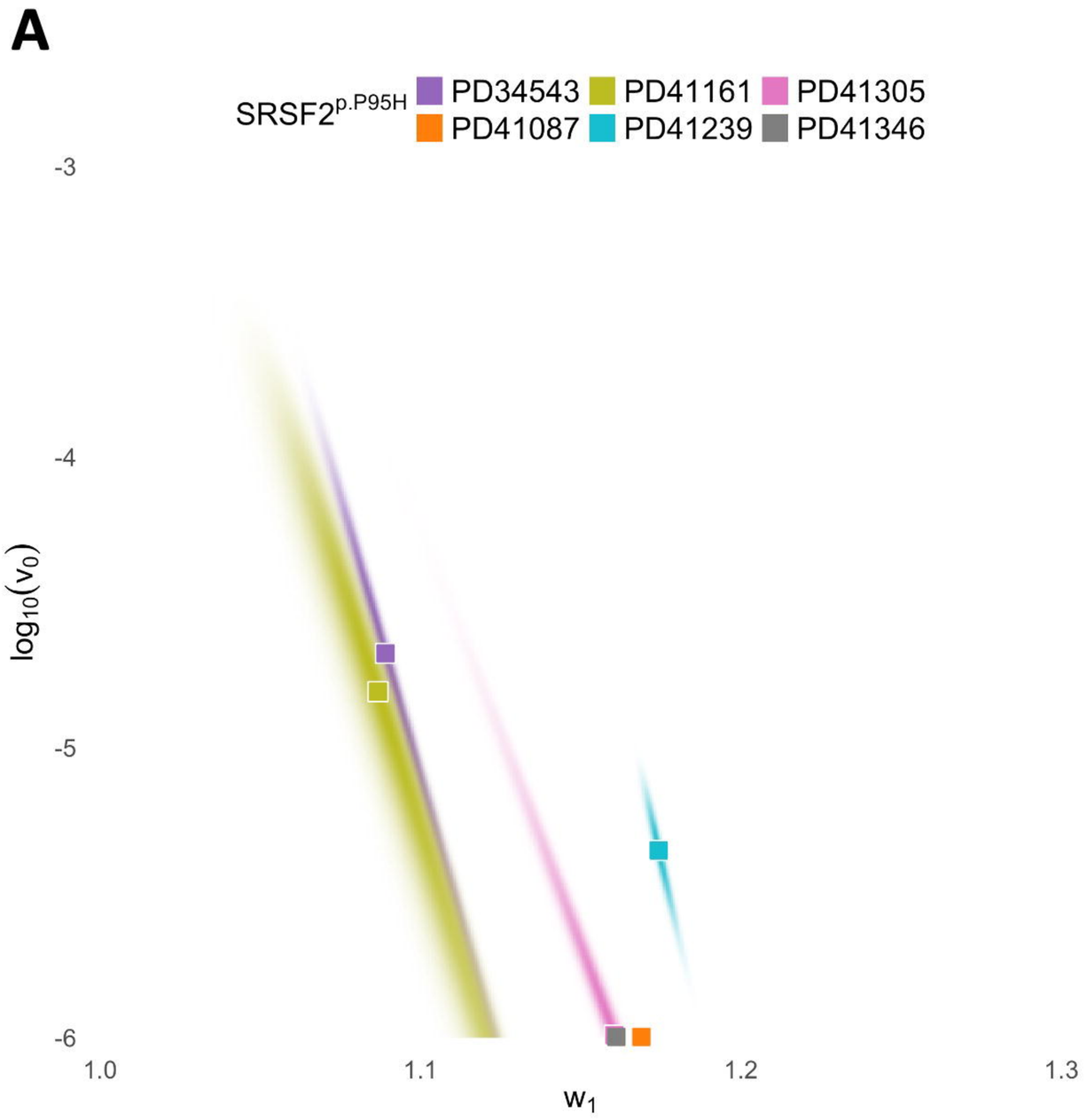

A
GNB1p.K57E
PD41082
PD41104
PD41132
PD41529
-3
-4
-5
-6
log10(v0)
1.0
1.1
1.2
1.3
w1
B
Variant Allele Frequency (VAF)
10−1
10−2
10−3
10−4
60
70
80
90
Age

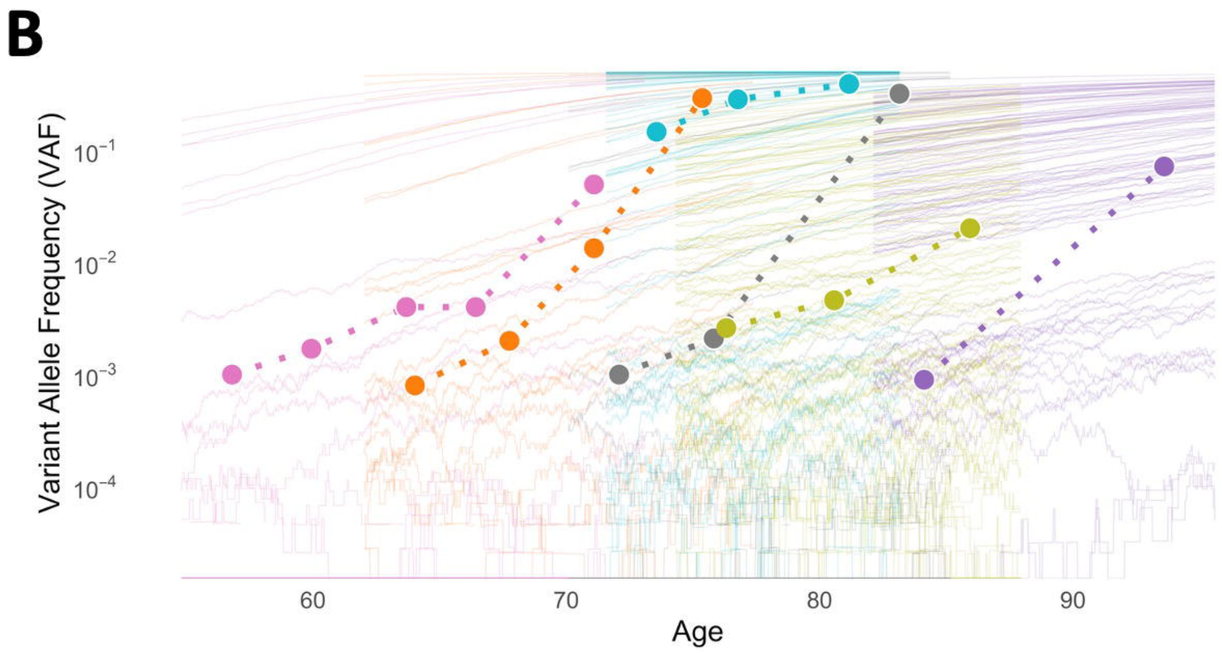

A
IDH2p.R140Q
PD41193
PD41214
PD41247
-3
-4
-5
-6
log10(v0)
1.0
1.1
1.2
1.3
w1

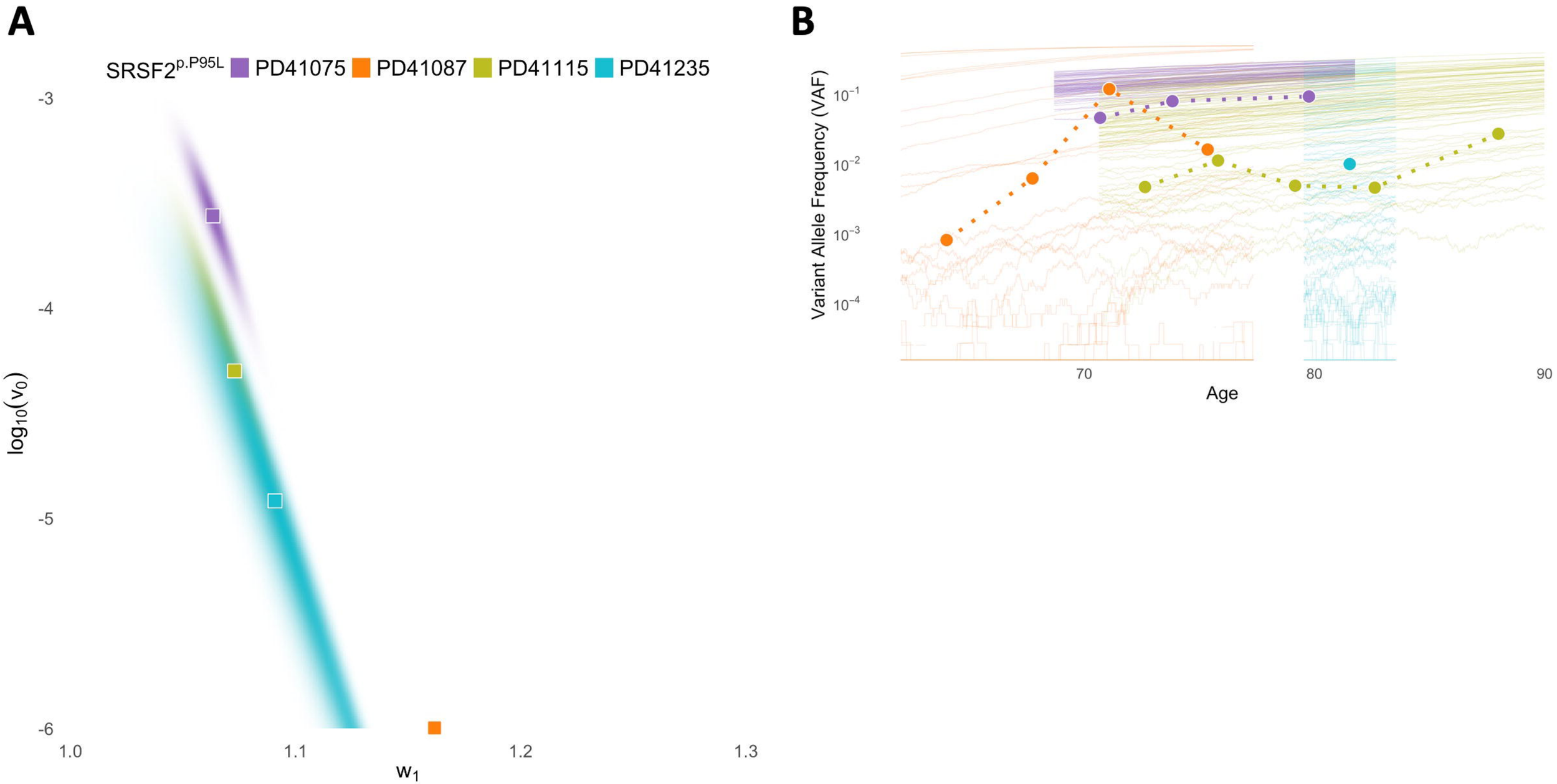

B
Variant Allele Frequency (VAF)
10−1
10−2
10−3
10−4
70
80
Age

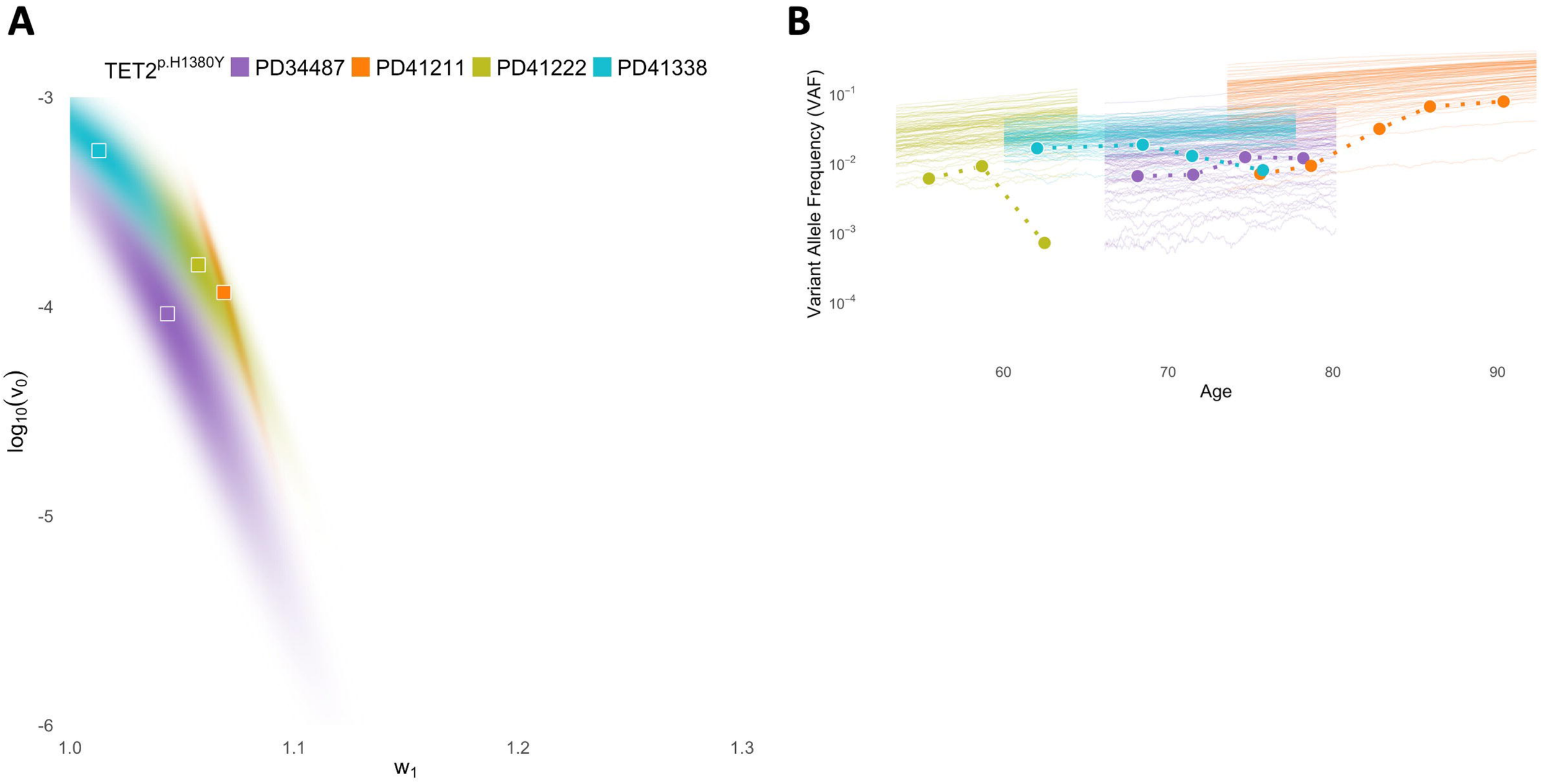

A
PPM1D^p.N477fs
PD34575
PD41253
PD41271
-3
-4
-5
-6
log10(v0)
1.0
1.1
1.2
1.3
w1

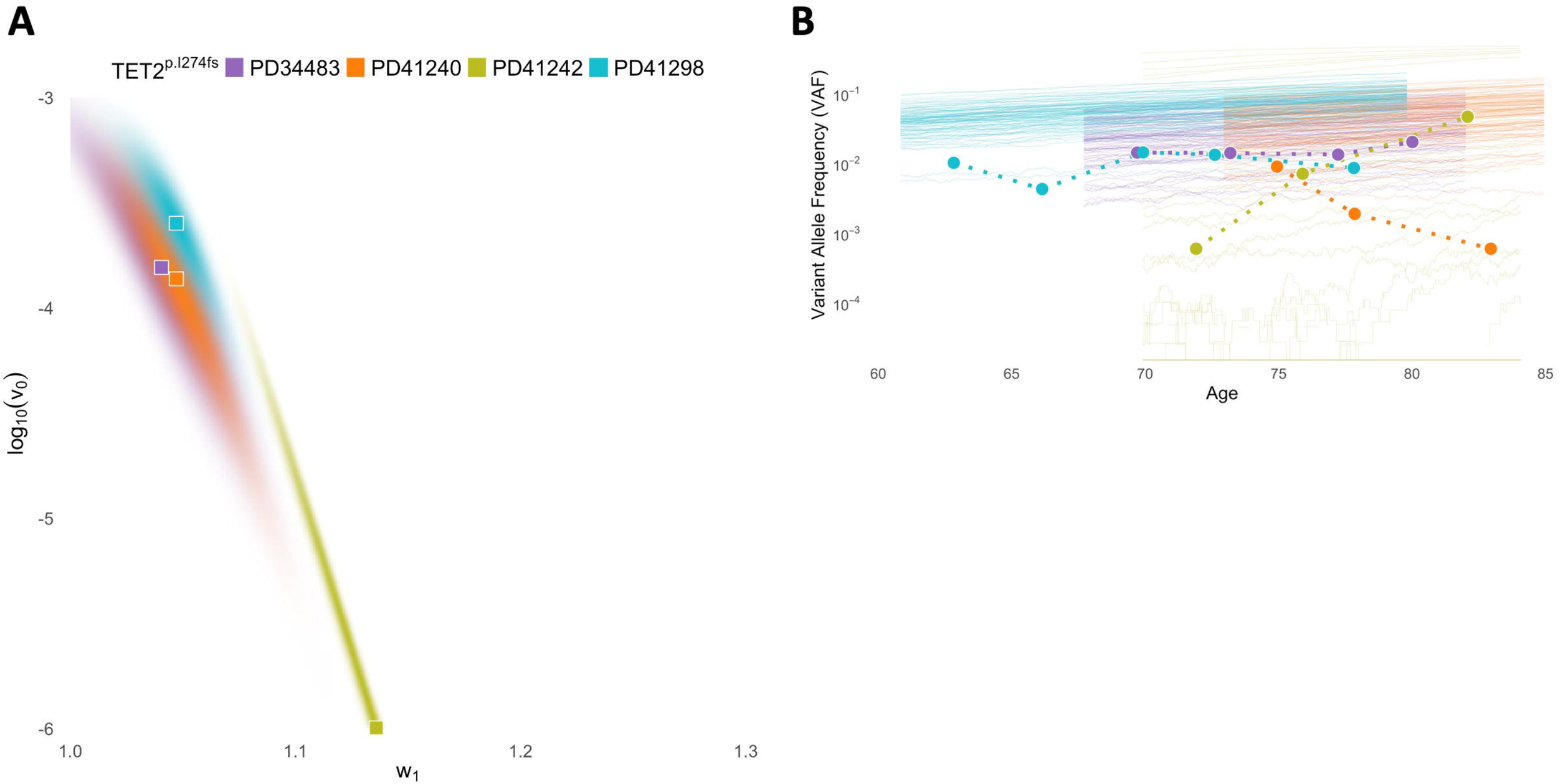

B
Variant Allele Frequency (VAF)
10−1
10−2
10−3
10−4
70
75
80
85
90
Age

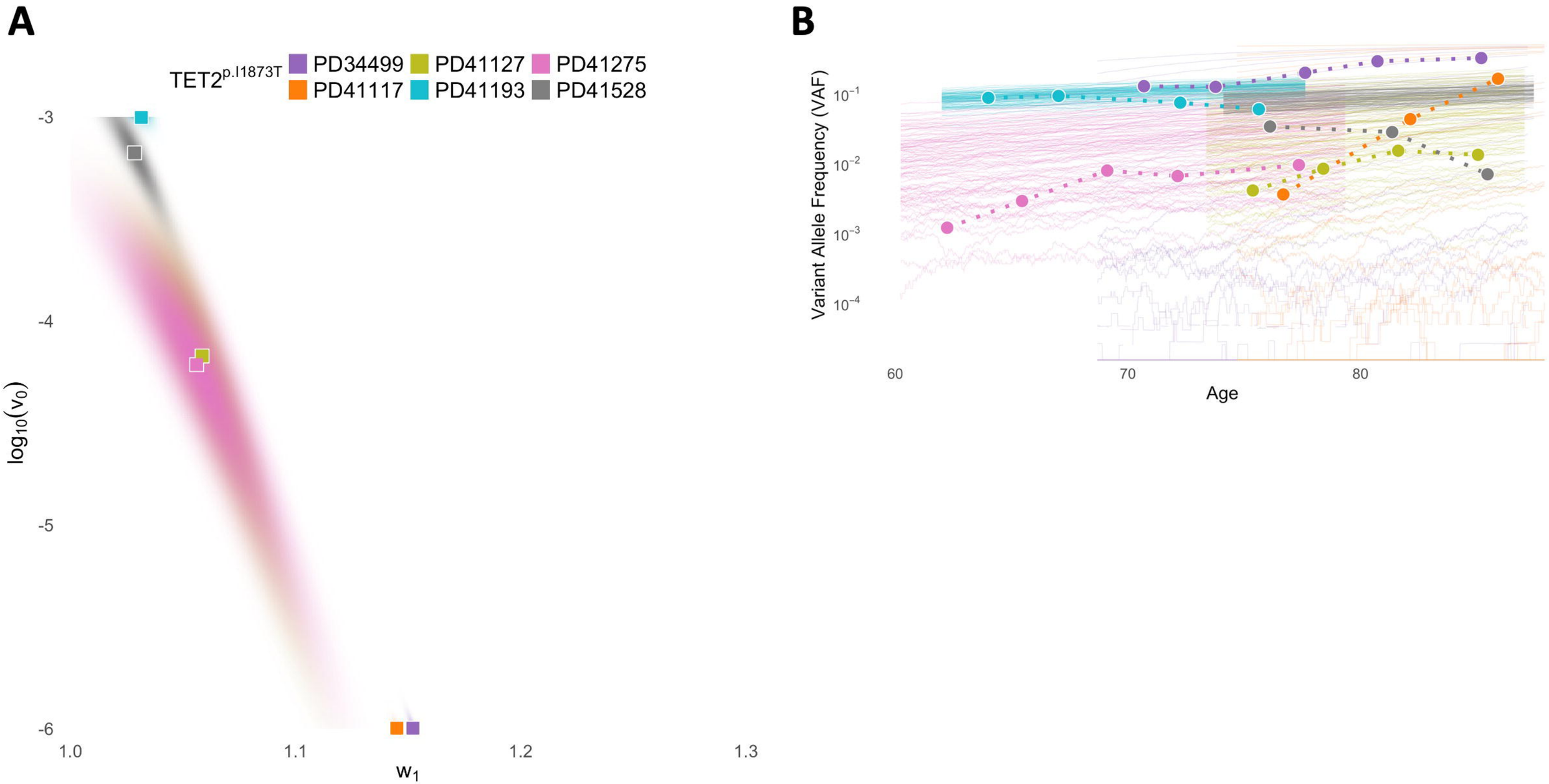

A
SRSF2p.P95H
PD34543
PD41161
PD41305
PD41087
PD41239
PD41346
-3
-4
-5
-6
log10(v0)
1.0
1.1
1.2
1.3
w1

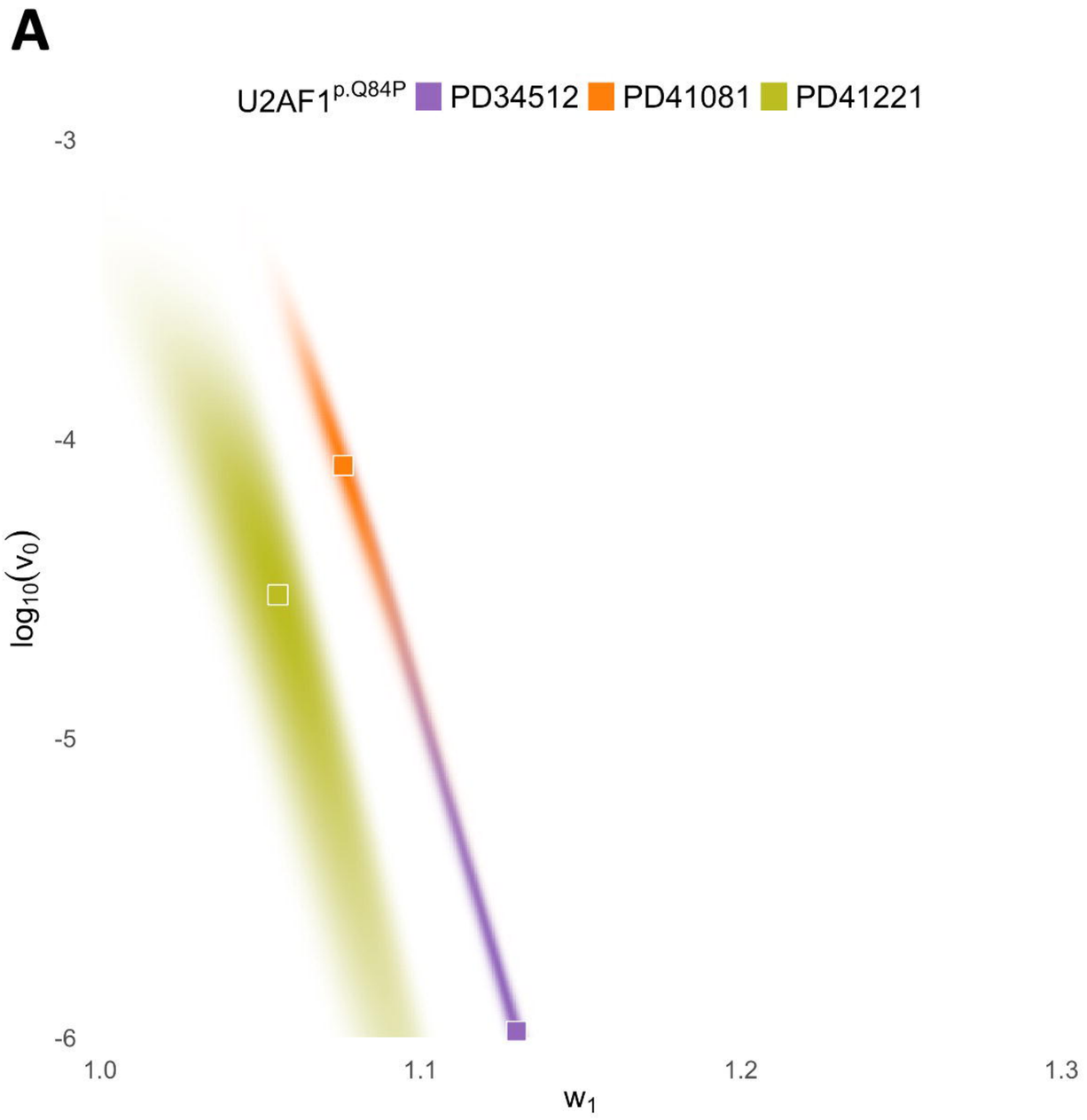

B
Variant Allele Frequency (VAF)
10−1
10−2
10−3
10−4
60
70
80
90
Age

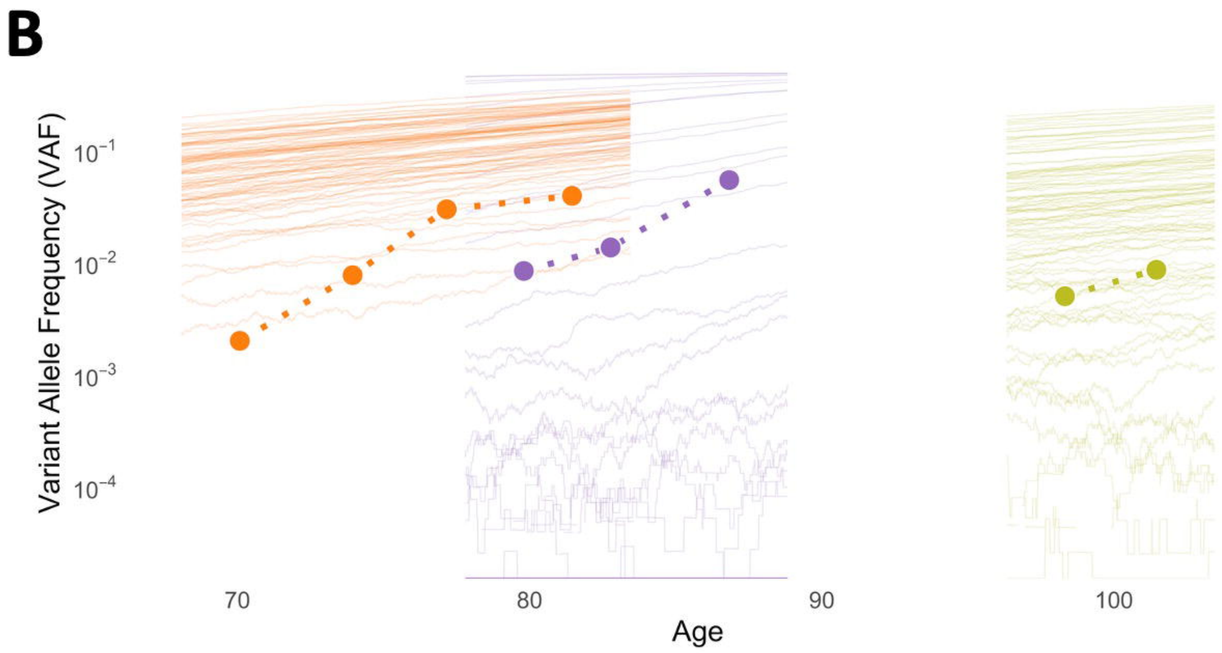

A
SRSF2p.P95L
PD41075
PD41087
PD41115
PD41235
-3
-4
-5
-6
log10(v0)
1.0
1.1
1.2
1.3
w1
B
Variant Allele Frequency (VAF)
10−1
10−2
10−3
10−4
70
80
90
Age


A
TET2p.H1380Y
PD34487
PD41211
PD41222
PD41338
-3
-4
-5
-6
log10(v0)
1.0
1.1
1.2
1.3
w1
B
Variant Allele Frequency (VAF)
10−1
10−2
10−3
10−4
60
70
80
90
Age


A
TET2p.I274fs PD34483 PD41240 PD41242 PD41298
-3
-4
-5
-6
log10(v0)
1.0
1.1
1.2
1.3
w1
B
Variant Allele Frequency (VAF)
10−1
10−2
10−3
10−4
60
65
70
75
80
85
Age


A
TET2p.I1873T
PD34499
PD41127
PD41275
PD41117
PD41193
PD41528
-3
-4
-5
-6
log10(v0)
1.0
1.1
1.2
1.3
w1
B
Variant Allele Frequency (VAF)
10−1
10−2
10−3
10−4
60
70
80
Age


A
U2AF1p.Q84P
PD34512
PD41081
PD41221
-3
-4
-5
-6
log10(v0)
1.0
1.1
1.2
1.3
w1


B
Variant Allele Frequency (VAF)
10−1
10−2
10−3
10−4
70
80
90
100
Age