# Inferring the initiation and development of myeloproliferative neoplasms

Gurvan Hermange, Alicia Rakotonirainy, Mahmoud Bentriou, Amandine Tisserand, Mira El-Khoury, François Girodon, Christophe Marzac, William Vainchenker, Isabelle Plo, Paul-Henry Cournède

# Inferring the initiation and development of Myeloproliferative Neoplasms


## Authors:

Gurvan Hermange[1], Alicia Rakotonirainy[1], Mahmoud Bentriou[1], Amandine Tisserand[2,3,4], Mira El-Khoury[2,3,4], François Girodon[6,7], Christophe Marzac[2,3,5,8], William Vainchenker[2,3,5], Isabelle Plo[2,3,5]*, Paul-Henry Cournède[1]*.

1 Université Paris-Saclay, CentraleSupélec, Laboratory of Mathematics and Informatics (MICS), Gif-sur-Yvette, France.

2- INSERM U1287 (INSERM, Gustave Roussy, Université Paris-Saclay), Villejuif, France

3- Gustave Roussy, Villejuif, France

4-Université de Paris (Paris Diderot), Paris, France

5- Université Paris-Saclay, Villejuif, France

6- Laboratoire d'Hématologie, CHU Dijon, Dijon, France

7- INSERM, UMR866, Centre de Recherche, Dijon, France

8- Laboratoire d'Immuno-Hématologie, Gustave Roussy, Villejuif

* corresponding authors: isabelle.plo@gustaveroussy.fr, paul-henry.cournede@centralesupelec.fr


# Abstract


The developmental history of blood cancer begins with mutation acquisition and the resulting malignant clone expansion. The two most prevalent driver mutations found in Myeloproliferative Neoplasms - *JAK2*$^{V617F}$ and *CALR*$^{m}$ - occur in hematopoietic stem cells, which are highly complex to observe *in vivo*. To circumvent this difficulty, we propose a method relying on mathematical modelling and statistical inference to determine disease initiation and dynamics. Our findings suggest that *CALR*$^{m}$ mutations tend to occur later in life than *JAK2*$^{V617F}$. Our results confirm the higher proliferative advantage of the *CALR*$^{m}$ malignant clone compared to *JAK2*$^{V617F}$. Furthermore,

we illustrate how mathematical modelling and Bayesian inference can be used for setting up early screening strategies.

## Significance

Myeloproliferative Neoplasms (MPN) blood cancers are often detected at an advanced age, after complications such as thrombosis and cardiovascular events. Understanding the MPN disease dynamics (initiation and development) could help set up early screening strategies, and more broadly, provide insight into the developmental history of blood cancer. For that purpose, we develop a mathematical approach combining modelling and statistical inference and apply it to the two most prevalent mutations in MPN. We evidence differences in the clonal expansion of the *CALR*$^{m}$ mutation compared to *JAK2*$^{V617F}$, suggesting different types of mechanisms for the driver mutation acquisition.

# Main

## Introduction

Myeloproliferative Neoplasms (MPNs) blood cancers are characterized by abnormal proliferation of myeloid hematological cells. These pathologies result from gain-of-function somatic mutations of specific genes occurring in hematopoietic stem cells (HSCs). Two main driver mutations of the MPN disease affect genes encoding proteins playing a crucial role in cell signaling: JAK2 (*JAK2*$^{V617F}$) and calreticulin (exon 9 mutations, *CALR*$^{m}$) (1–6). However, their precise role in the development of clonal hematopoiesis and in the disease dynamics remains poorly understood. Of particular debate is the expansion kinetics of a malignant clone emanating from a single mutated cell. While *JAK2*$^{V617F}$ was reported to be acquired decades before disease onset and even during fetal life (7–10), there is no such result for the *CALR*$^{m}$ mutation. MPNs are generally detected at advanced ages when the malignant clone is present at significant Variant Allele Frequency (VAF) in peripheral blood. Late detection, often after complications such as thrombosis and cardiovascular events, or more rarely at the leukemic stage, further increases the death rate. Methods to understand disease initiation and clonal expansion of mutated HSCs are thus crucial to develop adequate screening strategies and to avoid delayed medical care.

A major difficulty for this task comes from the fact that human HSCs are difficult to access and *in vivo* expansion of mutated HSCs cannot be observed. We therefore propose a mathematical model to infer the dynamics of the disease, from the acquisition of the driver mutation to the appearance of the symptoms. Stochastic models, such as the Wright-Fisher model used in

population genetics (11) or branching processes (12), are appropriate choices to describe the expansion of a clone from a unique mutated cell. However, their calibration from real observations remains challenging and requires setting up an extensive optimization procedure to achieve convergence of the algorithms in a realistic time.

To understand how *JAK2*$^{V617F}$ and *CALR*$^{m}$ malignant clones might expand over time, we propose a model relying on a Continuous Time Markov Chain (CTMC) process. When a certain number of mutated HSCs is reached, we show numerically that a deterministic law can approximate the process. This approximation further facilitates the calibration of the model that otherwise would not be computationally feasible. We estimate the model parameters using an Approximate Bayesian Computation method based on Sequential Monte Carlo (ABC-SMC) (13, 14), with data from recently diagnosed *JAK2*$^{V617F}$ and *CALR*$^{m}$ patients. To achieve a faster convergence, we use an optimal affectation procedure based on the Hungarian algorithm (15, 16). We apply this mathematical approach - combining modelling and statistical inference - to investigate potential differences in the disease dynamics of *CALR*$^{m}$ *vs JAK2*$^{V617F}$ MPN patients.

## Results

### *CALR*$^{m}$ mutations appear to be acquired later than *JAK2*$^{V617F}$

To analyze the initiation and dynamics in MPN disease and compare the expansion kinetics of *CALR*$^{m}$ *vs JAK2*$^{V617F}$ malignant clones, we propose an approach based on mathematical modelling (Fig. 1A) focusing on the proliferation dynamics of mutated HSCs (see Materials and Methods). In brief, we assume that a given mutated HSC divides at a rate $\alpha$ into 0, 1, or 2 mutated HSCs, depending on whether the division is differentiated, asymmetric, or symmetric, with probabilities $p_0$, $p_1$, and $p_2$, respectively (Fig. 1B). Daughter cells, if not HSCs, are progenitor cells that will further proliferate to give rise to mature cells (Fig. 1C). We denote by N(t) the number of mutated HSCs that have expanded from a single mutated cell at age $t > T_0$, with $T_0$ satisfying $N(T_0)=1$ (Fig. 1D). We further assume that the acquisition time $T_0$ follows an exponential distribution of mean $\lambda>0$ (Fig. 1D-iii). $T_0>0$ corresponds to an occurrence of the mutation after birth. Since Williams *et al.* reported that *JAK2*$^{V617F}$ might be acquired in fetal life (8), we also verify if our data could be in agreement with the hypothesis of mutation acquisition *in utero* ($T_0=0$) for all individuals, corresponding to the limit case $\lambda=0$ (Fig. 1D-ii).

Parameter values are estimated from single observations of 26 patients diagnosed with *JAK2*$^{V617F}$ or *CALR*$^{m}$ MPNs using a likelihood-free estimation procedure, namely ABC-SMC (17). Data ( D ), previously reported by Mosca *et al*. (18) and El-Khoury *et al.* (19), consist of Clonal Fractions (CF) for progenitor cells (see Materials and Methods and Fig. 2A). We estimate the model parameters separately for *JAK2*$^{V617F}$ and *CALR*$^{m}$ patient populations. The mean mutation acquisition time of the

*JAK2*$^{V617F}$ mutation E[λ] is estimated at ~15 years (Fig. 2B-i). Furthermore, concerning the hypothesis that mutation occurs in fetal life for the population of *JAK2*$^{V617F}$ patients, we estimate its probability p[ λ=0 |D] to be equal to 0.24. These findings are consistent with the values reported by Van Egeren *et al.* (7) and Williams *et al*. (8) (Supplemental Material E.1). Interestingly, for the *CALR*$^{m}$ patient population, we infer that almost surely λ>0, that is, the hypothesis of a mutation acquisition during fetal life for all *CALR*$^{m}$ patients is unlikely. We also estimate a higher expected mean acquisition time (E[λ]~25 years) than for *JAK2*$^{V617F}$ (Fig. 2B-ii and 2C-i).

**Higher proliferative advantage inferred for the *CALR*$^{m}$ malignant clone**

In the proposed model, the proliferative advantage of the mutated cells is described thanks to parameter $\Delta=p_2-p_0$, which represents the unbalance between symmetric and differentiated divisions that drives the clonal expansion of the mutation when Δ>0. The higher its value, the faster the malignant clone will expand. We estimate Δ for *CALR*$^{m}$ and *JAK2*$^{V617F}$ malignant clones separately and infer that the proliferative advantage of the *CALR*$^{m}$ malignant clone at the stem cell level is higher than that of *JAK2*$^{V617F}$ (Fig. 2B), with mean values of Δ respectively equal to 0.026 and 0.017. The propensity of *CALR*$^{m}$ HSCs for symmetric divisions and thus for the invasion of the stem cell pool is higher than that of *JAK2*$^{V617F}$ HSCs (Fig. 2C-ii). This observation explains that, even if the *CALR*$^{m}$ mutation is acquired later in life, the malignant clone can generally reach high CF more quickly (Fig. 2A). Our estimate of Δ for *JAK2*$^{V617F}$ is bigger than the one found by Watson *et al. (9)*, since we estimate Δ from observations of individuals having an MPN while Watson *et al*. studied clonal hematopoiesis of healthy donors (Supplemental Material E.2). Besides, we estimate a high probability of stochastic extinction ($q=p_0/p_2$) for *JAK2*$^{V617F}$ and *CALR*$^{m}$ malignant clones, with mean values of *q* respectively equal to 0.94 and 0.87, meaning that the acquisition of the mutation would lead to a clonal expansion - and then potentially a disease onset - in only ~10% of the cases (Supplemental Material D.2).

**Early screening might be a viable clinical option to detect *JAK2*$^{V617F}$ mutations**

Having estimated the mean parameter values of the model and their probability distributions, we can infer the MPN development and deduce early screening strategies (see Materials and Methods and Fig. 3A). The goal is to detect the mutation of interest in an individual as early as possible, i.e., before the onset of the disease. Early screening requires collecting blood samples from individuals and performing analysis of the gene of interest to measure the VAF in the peripheral blood (note that we always associate the term VAF with mature cells). The usual techniques in clinical practice can detect *CALR*$^{m}$ mutations at a VAF above 2% with CALR sizing and at around 0.1% for *JAK2*$^{V617F}$ (0.01% to 1% using allele-specific PCR). We consider that the detection is too late if the CF in HSCs is above 50% for *CALR*$^{m}$ (19) and above 15% for *JAK2*$^{V617F}$ (20), corresponding to the thresholds above which there is a high risk of MPN development and potential thrombosis. Our

approach leverages the posterior distribution of the model's parameters to compute the probability of detecting the mutation at different ages and thus it can help determine the best timing for screening. The optimal age for screening is found at 30 for the *JAK2*$^{V617F}$ mutation (Fig. 3B-i), and at 35 for the *CALR*$^{m}$ mutation (Fig. 3B-ii).

At this optimal age of detection, there are three possibilities for individuals having the mutation: the detection is too late because they already suffered from symptoms of the disease, we manage to detect the mutation early (true-positive), or we do not (false-negative). For *CALR*$^{m}$ patients, the probability of early detection - computed at the optimal screening age - remains low (42%), leaving 46% of the individuals who would later develop the disease undetected, and 12% for whom the detection would be too late. In contrast, early screening might be a viable clinical option to detect *JAK2*$^{V617F}$ mutations - with 79% of the individuals being detected early at the optimal screening age of 30.

Besides, we study how the sensitivity of the screening techniques influences the previous results. We compare different sensitivities with detection thresholds (corresponding to the VAF in mature cells) ranging from 0.01% to 2% (Fig 3C). Both for *CALR*$^{m}$ and *JAK2*$^{V617F}$ malignant clones, higher sensitivities increase the probability of detecting the mutation at lower ages. However, still at ultra-sensitive levels (detection threshold as low as 0.01%), the probability of early detection at the optimal screening age is below 65% for *CALR*$^{m}$, which is lower than the value obtained for *JAK2*$^{V617F}$ with a VAF threshold of 0.5% (Fig 3C-iii).

## Discussion

We developed a mathematical approach - combining modelling and statistical inference - to help infer the disease dynamics (initiation and development) and set up early screening strategies in MPNs, specifically *JAK2*$^{V617F}$ and *CALR*$^{m}$ MPNs. Both driver mutations occur in HSCs, making it difficult to study this disease experimentally in humans and justifies the use of mathematical models. These are necessarily simplifications of reality and are based on several assumptions, without which it would not be possible to make statistical inferences from data. We assumed that model parameters only depend on the type of the main driver mutation (*JAK2*$^{V617F}$ *vs CALR*$^{m}$) and not on the patients. Thus, heterogeneity between patients was only considered to result from stochastic effects. To refine our results and explore the influence of additional genetic factors, we would need to increase the number of patients included in our cohort to split them into subgroups of sufficient size. In particular, it would be relevant to increase the number of *CALR*$^{m}$ patients to account for differences between *CALR*$^{m}$ T1 and T2 mutations, which are clinically distinct (21). Besides, a limit of our work is that we only include patients without homozygous subclones. Yet, homozygous mutated cells are supposed to have an increased proliferative advantage that might

accelerate the disease development (8). Thus, an extension of our work would be to complexify the model by integrating the homozygous subclones dynamics. Such an extended model could also describe the expansion of subclones with additional MPN mutations (*TET2*, *DNMT3A,…)*. The proliferative advantage of malignant clones, modeled by Δ, was also considered constant over life. However, more accurate models should probably take into account threshold effects, possibly justified by regulatory mechanisms, to avoid the possibility of unrealistic high clonal expansion. WT HSCs are also likely to acquire mutations that increase their proliferative advantage, reducing the relative selective advantage of $JAK2^{V617F}$ or $CALR^{m}$ clones, while these latter could also acquire associated mutations, increasing their proliferative advantage. Parameter Δ should be thus considered an averaged value that embeds different effects occurring over life. Besides, we implicitly considered that the evolution of MPN blood cancers is driven by natural selection by assuming Δ>0 and proliferation of the malignant clone occurring in a large and constant pool of WT HSC. As discussed in Lyne *et al*. (22), a proper estimation of the number of WT HSCs would be essential to distinguish natural selection from neutral evolution, but also in our case, to quantify more precisely parameter Δ. Furthermore, the fact that aged HSCs might encounter more symmetric divisions, as discussed in Florian *et al*. would imply that the number of WT HSC increases over age and does not stay constant (23). More complex models could be developed, but they would require longitudinal observations and larger datasets for their calibration. Our model has the advantage that it can be used to study other mutations occurring in various blood cancers and that its calibration only needs measures of CF among progenitor cells. Applied to the $JAK2^{V617F}$ mutation, we found a mean acquisition time consistent with other studies (7, 8). Our estimate of parameter Δ is bigger than that obtained by Watson et al., who studied clonal hematopoiesis of normal individuals (9); the aggressiveness of MPN appears compatible with a higher proliferative advantage of the mutant clones compared to general clonal hematopoiesis (24–26). Furthermore, by applying the same approach to $CALR^{m}$ patients, we could evidence differences between the clonal expansion of the two main driver mutations in MPN. These different effects of $JAK2^{V617F}$ and $CALR^{m}$ have also been observed in mouse models through competitive engraftment of bone marrow cells at limiting dilution of HSC in which the $CALR^{m}$ HSC outcompete WT cells more rapidly than $JAK2^{V617F}$ HSCs to induce the disease (27–29). In agreement with our data, it has been shown in early progenitors from essential thrombocythemia patients that $JAK2^{V617F}$ gave a lower clonal dominance than $CALR^{m}$ (7, 19, 20, 30). Moreover, our results suggest that the $CALR^{m}$ mutation is unlikely acquired during fetal life, but on average 25 years after birth, in contrast to the $JAK2^{V617F}$ mutation. It is consistent with observations that young MPN patients with inaugural Budd-Chiari syndrome (mean age of 35 years) harbor more $JAK2^{V617F}$ than $CALR^{m}$ (90% versus 2%) (31). Yet, recent findings have also evidenced a possible acquisition of CALRm mutation *in utero* (32). Different types of mechanisms seem to exist for the driver mutation acquisition. A possible explanation is that a difference in the type of mutation, i.e., a point mutation for $JAK2^{V617F}$ and

deletions/insertions with +1 frameshift for *CALR* mutations, could result in two distinct types of deficient DNA repair mechanisms (leading to G->T transversion and to deletion/insertion), and that these defects might occur at different life ages. An interesting observation from Cordua *et al*. (33) is that even if $JAK2^{V617F}$ is more frequent than $CALR^{m}$ mutation in the general healthy population, the proportion of diseased patients in $CALR^{m}$ individuals is much higher than in $JAK2^{V617F}$ individuals. It suggests that the latency of the disease is shorter with the $CALR^{m}$ mutation and that the $CALR^{m}$ clone expands faster than $JAK2^{V617F}$, in keeping with our results. Moreover, even if the $JAK2^{V617F}$ mutation occurs before $CALR^{m}$, the faster expansion for $CALR^{m}$ can explain that the $CALR^{m}$ disease starts 10 years in average before the $JAK2^{V617F}$ disease (34, 35).

Finally, our inference method relies on a Bayesian framework providing the parameter probability distributions, and thus a proper assessment of parameter uncertainty. Such an approach is computationally expensive and relies on efficient numerical methods to obtain convergence of the algorithms in a realistic time. We set up an optimization procedure to make this study possible, deriving a deterministic approximation of our stochastic model, and using an optimal assignment algorithm. Then, from the posterior distributions of the parameters, we were able to infer more precisely the dynamics of the clonal expansion and explore strategies for early screening. We found that almost half of $CALR^{m}$ patients would not be detected at the optimal screening age. Therefore, testing for the *CALR* mutation is not recommended unless the individuals are examined several times in their lives. In contrast, we found that early detection might be a viable clinical option to detect $JAK2^{V617F}$ mutation, with an optimal age for measuring VAF in peripheral blood of around 30 years. Since this mutation is present in a substantial number of cases with unexplained splanchnic and cerebral thromboses for patients that exhibit a non-diagnosed MPN with low $JAK2^{V617F}$ VAF, early detection may prevent thrombotic events.

To conclude, our method - applied to the two most prevalent driver mutations for MPNs - illustrates the potential of mathematical modelling to help infer the timing of blood cancer initiation, better understand its development, and design early detection strategies tailored to the type of mutation. In the future, it may be possible to test for many malignant mutations through NGS simultaneously instead of considering each of them separately. Such a large-scale screening effort would result in a more complex optimization problem; yet, the flexibility of our modelling framework makes its extension to a panel of genes feasible. In the future, generalizing the application of our mathematical approach could yield pragmatic and efficient prevention strategies against blood cancers.

## Materials and Methods

### Model and parameters

To model the MPN disease dynamics, we consider a clonal expansion in a stem cell pool whose number $N_{WT}$ of WT cells is assumed constant and equal to 100,000 (36). The expansion from a single mutated HSC begins at age $T_0$. *N(t)* is the number of mutated cells at age *t*. The invasion of the stem cell pool is modeled by considering an unbalance between symmetric (occurring with probability $p_2$) and differentiated divisions (occurring with probability $p_0$). $\Delta=p_2-p_0$, is the parameter that describes the proliferative advantage of the malignant clone, as introduced by Mosca *et al*. (18). Considering that the mutation confers a proliferative advantage at the stem cell level, we have $\Delta>0$. We further assume that a given mutated HSC divides at a rate α. Mathematically, this expansion process is described as a CTMC, where the probability of extinction *q* of the mutant cell population is equal to $q=p_0/p_2$. Throughout all the article, we focus on trajectories without extinction. During the expansion process, mutated HSCs also give rise to progenitor cells. Although these latter arise from HSCs, their CF is not a proxy for the CF among HSCs. However, by model identification and notably estimating parameter Δ, we can infer the CF in the progenitor compartment at age *t*, denoted by $\eta(t)$:

$$\eta(t) = \frac{(1-\Delta)N(t)}{(1-\Delta)N(t)+N_{WT}} \quad (1)$$

Progenitor cells then give rise to mature cells. In clinical routine, the VAF measured in peripheral blood provides insight into the disease progression (note that we only consider heterozygous mutated cells: therefore, the percentage of mutated cells is twice that of mutated alleles). *JAK2*$^{V617F}$ but also, to a lesser extent, *CALR*$^{m}$ progenitor cells proliferate more than WT cells during the latest stages of hematopoiesis. Extending eq.(1), the VAF in peripheral blood at time t is equal to:

$$\mathrm{VAF}(t) = 0.5\frac{(1-\Delta)k_m N(t)}{(1-\Delta)k_m N(t)+N_{WT}} \quad (2)$$

where $k_m$ models the proliferative advantage of mutated cells, from progenitors to mature cells, compared to WT cells.

We consider an acquisition time $T_0$ that follows an exponential distribution. Without prior knowledge, the exponential law is an appropriate and convenient choice that introduces only one additional parameter to estimate, namely the mean of the exponential distribution denoted by $\lambda>0$. $T_0 = 0$ corresponds in our model to an acquisition in fetal life; it is not a punctual event but rather a period of about 0.75 years during which the WT stem cell pool quickly expands. Williams *et al.* reported that *JAK2*$^{V617F}$ mutation could be acquired during this period (8). Therefore, we study two hypotheses: either the mutation occurs (for all patients of one of our two populations of interest, i.e., with *CALR*$^{m}$ or with *JAK2*$^{V617F}$) in fetal life, that is, $T_0=0$, or later after birth, $T_0>0$. We consider

that the latter corresponds to λ>0 when the former corresponds to the limit case λ=0, the problem of model selection (selection between both hypotheses) is thus reduced to a question of parameter estimation. We consider *a priori* that λ=0 with a non-zero probability, such that its posterior could be a zero-inflated distribution. p[ λ=0 | D] will correspond to the probability of mutation acquisition in fetal life (Supplemental Material B.4).

More details about the model, its assumptions, and the prior distributions of the parameters are given in Supplemental Material A.

**Data**

Parameter values are estimated from the observations of 15 patients diagnosed with *CALR*$^{m}$ MPNs (19), and 11 patients diagnosed with *JAK2*$^{V617F}$ MPNs (18). Data were obtained by determining the clonal architecture of the CD34$^{+}$ progenitors purified from blood samples (Tables D1-2 in Supplemental Material). Thus, we get for each MPN patient *i* the CF of progenitor cells $\eta(t_i)$ at time $t_i$ that can be confronted with the theoretical value given by eq. (1). We only consider patients with heterozygous clones. Some patients have associated mutations with clonal hematopoiesis, so the patients of our cohort can be regarded as representative of the MPN population. Our model does not directly account for differences between patients, except for the main driver mutation (*CALR*$^{m}$ or *JAK2*$^{V617F}$).

More details are given in Supplemental Material D.1.

**Likelihood-free estimation procedure**

We consider a Bayesian framework and estimate the parameter vector's posterior distribution given the observed data *D*: $p[\theta|D] \propto p[D|\theta]p[\theta]$ (where $\theta$ is the parameter vector).

For CTMC, the expression of the likelihood for pointwise observations is generally intractable. We, therefore, use a likelihood-free estimation procedure, namely ABC-SMC (17). With this method, the posterior distribution is obtained by iterative rejection sampling, using a decreasing sequence of tolerances (Supplemental Material B.1). In brief, the posterior distribution will be approximated based on M particles (we chose M=2,000), that is, M parameter vectors. For a given particle, a parameter vector is sampled (initially from the prior distribution), the model is simulated, and generates as many dynamics as we have observations. The quadratic error (distance *d*) between data and simulations is computed. If *d* is below the considered tolerance (at the current stage of the algorithm), the parameter vector is assigned to the particle. If not, the previous steps are repeated. When each particle is associated with a parameter value, the method is repeated with a lower tolerance until the algorithm has converged, that is, until the estimation of the posterior distribution no longer changes when decreasing the tolerance (Fig. B1-3).

ABC-SMC relies on extensive model simulations and their confrontation to experimental observations in our patients' populations (first *JAK2*$^{V617F}$, then *CALR*$^{m}$). To achieve computation within a reasonable time (less than 6 hours when running the computations in parallel using 250 processors of an HPC cluster), we optimized the procedure. First, we derived a hybrid model, considering that only the early invasion of the stem cell pool by the malignant clone is driven by stochastic effects and then switching to a deterministic expansion when we reach a sufficient number of mutated HSCs $N_c$=2,000 (i.e., a CF approximately equal to 2%). The deterministic expansion is an approximation valid for $N_c$ large. The choice of $N_c$ - determined through a numerical calibration procedure - is a compromise between reducing the approximation error and increasing the computation speed, that impacts the quality of the inference (Supplemental Material B.2, Fig. B4-6). Second, we use an optimal assignment procedure for ABC-SMC, the Hungarian algorithm (15, 16) when comparing the simulations of all individuals to the patients' observations (Supplemental Material B.3).

Using synthetic datasets, we check that the number of patient observations is sufficient to draw robust conclusions from the proposed parameter estimation procedure (Supplemental Material C.1-2, Tab. C1). We study how different model assumptions impact the estimation of parameter Δ (Supplemental Material C.3). We then run these computations on the actual dataset, and evaluate the model fit quality through a leave-one-out analysis (Supplemental Material D). Finally, we assess to what extent our results for *JAK2*$^{V617F}$ were consistent with other reports (Supplemental Material E).

**Optimal strategies for early screening**

Having estimated the parameter posterior distributions, we can infer the dynamics of the clonal expansion and thus deduce optimal strategies for early screening. We face an optimization problem. If too early, the screening will come with an essential risk of false-negative results; a high proportion of people would develop the disease but not be identified. On the other hand, if too late, people might be detected when their stem cell pool is already invaded. With the proposed model, we can then solve this optimization problem numerically since we infer the dynamics of the malignant clone initiation and expansion. We consider that the detection is too late when the CF among HSCs is above a threshold $CF_{late}$:

$$N(t)/(N(t)+N_{WT}) > CF_{late}$$

with $CF_{late}$ =50% for *CALR*$^{m}$ (19) and equal to 15% for *JAK2*$^{V617F}$ (20), corresponding to the thresholds above which there is a risk of MPN onset and potential thrombosis.

In clinical routine, only the VAF in peripheral blood is measured. However, the VAF measure is not a good proxy for the CF among progenitors, let alone stem cells. In our model, VAF(t) at time t is given by eq. (2). We consider that an existing mutation is not detected (false-negative) when VAF(t)< $VAF_{detection}$ where $VAF_{detection}$=0.1% for *JAK2*$^{V617F}$ and $VAF_{detection}$=2% for *CALR*$^{m}$.

Finally, the optimization problem to solve is to find the age at which the probability of early screening in the population is as high as possible:

$$T^* = \max_t \mathbb{P}\left[\mathrm{VAF}(t) > \mathrm{VAF}_{detection}\ ;\ N(t)/(N(t)+N_{WT}) < CF_{late} | \mathcal{D}\right]$$

We solve this problem numerically for *JAK2*$^{V617F}$ and *CALR*$^{m}$ separately, evaluating the probability of early detection at age *t* by sampling parameters from the estimated posterior distribution. We simulate 20,000 trajectories this way and represent in figure 3B the probability of early screening as well as the rate of false-negative results and the probability of a too late detection (MPN symptoms), from which we can deduce the optimal age for screening.

Then, we study the effect of the sensitivity of the techniques that detect the mutation on these results (Fig. 3C).

## Data and code availability

Data were previously reported by Mosca *et al*. (18) and El-Khoury *et al*. (19) (Supplemental Material Tab. D1-2).

The ABC-SMC framework used to calibrate the model was implemented using the programming language Julia and made available on GitLab:

https://gitlab-research.centralesupelec.fr/2017bentrioum/markovprocesses.jl

as well as the model implementation:

https://gitlab-research.centralesupelec.fr/2012hermangeg/mpn-development

# References


1. R. L. Levine, *et al.*, Activating mutation in the tyrosine kinase JAK2 in polycythemia vera, essential thrombocythemia, and myeloid metaplasia with myelofibrosis. *Cancer Cell* **7**, 387–97 (2005).

2. C. James, *et al.*, A unique clonal JAK2 mutation leading to constitutive signalling causes polycythaemia vera. *Nature* **434**, 1144–8 (2005).

3. E. J. Baxter, *et al.*, Acquired mutation of the tyrosine kinase JAK2 in human myeloproliferative disorders. *Lancet* **365**, 1054–61 (2005).

4. R. Kralovics, *et al.*, A gain-of-function mutation of JAK2 in myeloproliferative disorders. *N Engl J Med* **352**, 1779–90 (2005).

5. J. Nangalia, *et al.*, Somatic CALR mutations in myeloproliferative neoplasms with nonmutated JAK2. *N Engl J Med* **369**, 2391–405 (2013).

6. T. Klampfl, *et al.*, Somatic mutations of calreticulin in myeloproliferative neoplasms. *N Engl J Med* **369**, 2379–90 (2013).

7. D. Van Egeren, *et al.*, Reconstructing the Lineage Histories and Differentiation Trajectories of Individual Cancer Cells in Myeloproliferative Neoplasms. *Cell Stem Cell* (2021) https:/doi.org/10.1016/j.stem.2021.02.001.

8. N. Williams, *et al.*, Life histories of myeloproliferative neoplasms inferred from phylogenies. *Nature* **602**, 162–168 (2022).

9. C. J. Watson, *et al.*, The evolutionary dynamics and fitness landscape of clonal hematopoiesis. *Science* **367**, 1449–1454 (2020).

10. P. Hirsch, *et al.*, Clonal history of a cord blood donor cell leukemia with prenatal somatic JAK2 V617F mutation. *Leukemia* **30**, 1756–9 (2016).

11. W. J. Ewens, *Mathematical population genetics: theoretical introduction.*, Springer, New York (2004).

12. M. Kimmel, D. E. Axelrod, *Branching Processes in Biology*, Springer, New York, NY (2015).

13. S. A. Sisson, Y. Fan, M. M. Tanaka, Sequential Monte Carlo without likelihoods. *Proc Natl Acad Sci U S A* **104**, 1760–1765 (2007).

14. T. Toni, D. Welch, N. Strelkowa, A. Ipsen, M. P. H. Stumpf, Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J R Soc Interface* **6**, 187–202 (2009).

15. H. W. Kuhn, The Hungarian method for the assignment problem. *Naval Research Logistics* **2**, 83–97 (1955).

16. J. Munkres, Algorithms for the Assignment and Transportation Problems. *Journal of the Society for Industrial and Applied Mathematics* **5**, 32–38 (1957).

17. M. A. Beaumont, J.-M. Cornuet, J.-M. Marin, C. P. Robert, Adaptive approximate Bayesian computation. *Biometrika* **96**, 983–990 (2009).

18. M. Mosca, *et al.*, Inferring the dynamic of mutated hematopoietic stem and progenitor cells induced by IFNα in myeloproliferative neoplasms. *Blood*, blood.2021010986 (2021).

19. M. El-Khoury, *et al.*, Different impact of calreticulin mutations on human hematopoiesis in myeloproliferative neoplasms. *Oncogene* **39**, 5323–5337 (2020).

20. S. Dupont, *et al.*, The JAK2 617V>F mutation triggers erythropoietin hypersensitivity and terminal erythroid amplification in primary cells from patients with polycythemia vera. *Blood* **110**, 1013–21 (2007).

21. A. Tefferi, *et al.*, Type 1 versus Type 2 calreticulin mutations in essential thrombocythemia: a collaborative study of 1027 patients. *Am J Hematol* **89**, E121-4 (2014).

22. A.-M. Lyne, L. Laplane, L. Perié, To portray clonal evolution in blood cancer, count your stem cells. *Blood* **137**, 1862–1870 (2021).

23. M. C. Florian, *et al.*, Aging alters the epigenetic asymmetry of HSC division. *PLoS Biol* **16**, e2003389 (2018).

24. S. Jaiswal, *et al.*, Age-related clonal hematopoiesis associated with adverse outcomes. *N Engl J Med* **371**, 2488–98 (2014).

25. G. Genovese, *et al.*, Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N Engl J Med* **371**, 2477–87 (2014).

26. T. McKerrell, *et al.*, Leukemia-associated somatic mutations drive distinct patterns of age-related clonal hemopoiesis. *Cell Rep* **10**, 1239–45 (2015).

27. S. Hasan, *et al.*, JAK2V617F expression in mice amplifies early hematopoietic cells and gives them a competitive advantage that is hampered by IFNalpha. *Blood* **122**, 1464–77 (2013).

28. C. Benlabiod, *et al.*, Calreticulin del52 and ins5 knock-in mice recapitulate different myeloproliferative phenotypes observed in patients with MPN. *Nat Commun* **11**, 4886 (2020).

29. P. Lundberg, *et al.*, Myeloproliferative neoplasms can be initiated from a single hematopoietic stem cell expressing JAK2-V617F. *J Exp Med* **211**, 2213–30 (2014).

30. S. Anand, *et al.*, Effects of the JAK2 mutation on the hematopoietic stem and progenitor compartment in human myeloproliferative neoplasms. *Blood* **118**, 177–81 (2011).

31. J. Poisson, *et al.*, Selective testing for calreticulin gene mutations in patients with splanchnic vein thrombosis: A prospective cohort study. *Journal of Hepatology* **67**, 501–507 (2017).

32. N. Sousos, *et al.*, In utero origin of myelofibrosis presenting in adult monozygotic twins. *Nat Med* (2022) https:/doi.org/10.1038/s41591-022-01793-4.

33. S. Cordua, *et al.*, Prevalence and phenotypes of JAK2 V617F and calreticulin mutations in a Danish general population. *Blood* **134**, 469–479 (2019).

34. A. Tefferi, J. Thiele, A. M. Vannucchi, T. Barbui, An overview on CALR and CSF3R mutations and a proposal for revision of WHO diagnostic criteria for myeloproliferative neoplasms. *Leukemia* **28**, 1407–13 (2014).

35. E. Rumi, *et al.*, JAK2 or CALR mutation status defines subtypes of essential

thrombocythemia with substantially different clinical course and outcomes. *Blood* **123**, 1544–51 (2014).

36. H. Lee-Six, *et al.*, Population dynamics of normal human blood inferred from somatic mutations. *Nature* **561**, 473–478 (2018).

# Figure Legends

**Fig. 1: Method and model for inferring the development of MPN**

A. Overview of the method.

B. Modelling the stem cell proliferation dynamics: a given mutated HSC divides after a random time T that follows an exponential law ε(α) of rate α. After dividing, the mutated HSC gives birth to 0, 1, or 2 mutated HSC with probabilities $p_0$, $p_1$, or $p_2$, respectively. Daughter cells, if not HSCs, are progenitor cells (represented in a darker color). On this example, the first division is symmetric, and the division of one of the daughter cell (on the top) is asymmetric.

C. The hematopoietic dynamics can be modeled with three compartments: i) the stem cell pool consisting of a limited number of HSCs that can self-renew and thus drive the long-term clonal expansion; ii) the progenitor cells ($CD34^+$) that arise from HSCs. Our model of MPN expansion is calibrated from measured CF of those cells; iii) a pool of mature cells, whose measure of the VAF in clinical routine can allow the early detection of the mutation. Blue cells are mutated.

D. Modelling the stochastic expansion of the malignant clone from a single mutated cell, where N(t) is the number of mutated HSCs over time t. Two hypotheses are compared for the modelling of the mutation acquisition: either i) after birth ($T_0>0$) or ii) during fetal life ($T_0=0$). In the former case, $T_0$ follows an exponential law ε(1/λ), of rate 1/λ (where λ>0 corresponds to the mean acquisition time), whose probability density function (pdf) is represented in (iii). For each scenario 100 simulations are computed, with α = 1/30, Δ = 0.02, q =1/3, and λ=10 (for (i) only).

**Fig. 2: Inferring the MPN development**

A. Confronting *JAK2*$^{V617F}$ (i) or *CALR*$^{m}$ (ii) patient observations (black points) to the inferred evolution of the CF among progenitor cells. The dynamics of clonal expansion over time are obtained from 10,000 simulations of the model, sampling the parameters from the estimated posterior distributions. The bold line represents the median CF, computed over 10,000 simulated trajectories at each age, when the lines on both sides materialize 25, 50, 75, 90, 95 and 99%

credibility intervals. The gradient of color indicates where the trajectories can be found; the darker, the higher probability.

B. Estimated joint posterior distributions of parameters Δ and λ, for *JAK2*$^{V617F}$ (i) or *CALR*$^{m}$ (ii). Black solid lines in the 1D histograms correspond to the mean values. The darkest regions on the 2D histograms correspond to those of higher probability density.

C. Schematic representation of the MPN development for *JAK2*$^{V617F}$ and *CALR*$^{m}$ malignant clones.

i) *JAK2*$^{V617F}$ mutation is found to be acquired earlier on average than *CALR*$^{m}$ mutation (15 vs 25 years old), and potentially during fetal life (with probability p[λ=0|D] = 0.24 when the same probability is estimated at zero for *CALR*$^{m}$). The histograms represent the distribution of parameter λ.

ii) *CALR*$^{m}$ is found to have a higher proliferative advantage at the stem cell level than *JAK2*$^{V617F}$. *CALR*$^{m}$ malignant clone (in orange) will expand over time - in a pool of wild-type (WT) HSCs (purple cells) - at a higher rate than *JAK2*$^{V617F}$ (in blue).

**Fig. 3: Early screening**

A. Overview of the method for determining the optimal screening age.

i) One trajectory of the stochastic model (after its calibration from actual observations) corresponds to the disease expansion in an arbitrary (randomly sampled) patient. From the model, we obtain the progression of the VAF in mature cells over the years. Three periods in the patient's life can be considered. First, the VAF is lower than a detection threshold (grey area); the malignant clone has already begun expanding but is still undetectable. Then (green area), the mutation becomes detectable still with a sufficiently low CF so that there is a low risk of MPN symptoms. Theoretically, it would be the appropriate period for the early screening of this particular patient. Eventually, the malignant clone continues to expand; the VAF exceeds a threshold above which there is a risk of MPN symptoms (red area): screening would be too late.

ii) Because of the heterogeneity between patients, the three previously defined periods (mutation not detected - grey, early detection - green, and MPN symptoms - red) are different between individuals.

iii) Considering a high number of MPN patients, we obtain as functions of age the frequencies (or probabilities) of early screening (green line), to not detect the mutation (dashed grey line), or to have MPN symptoms (red dashed line). The probability of early screening reaches a maximum at the optimal screening age, that is, the age at which it would be optimal to test the population for the

considered mutation. The value reached at the optimal screening age corresponds to the highest proportion of patients that would be detected early enough (according to our mathematical model and its calibration). The higher this value, the more efficient the screening.

B. Evolution of the probability of early detection (solid line, blue for *JAK2*$^{V617F}$ (i) and orange for *CALR*$^{m}$ (ii)), false negative rate (dashed line), and too late detection (dash-dotted line) when testing the population for the *JAK2*$^{V617F}$ (i) or *CALR*$^{m}$ (ii) mutation at different ages. Here, we consider that *CALR*$^{m}$ mutation is detected in mature cells for a VAF higher than 2% (sizing CALR) and that *JAK2*$^{V617F}$ is detected for a VAF higher than 0.1% (allele-specific PCR).

C. Impact of the sensitivity of the techniques for detecting *JAK2*$^{V617F}$ (i) or *CALR*$^{m}$ (ii) mutations. As expected, higher sensitivities increase the probability to detect the mutations, at earlier ages. In (iii), we compute the maximal probability to detect the mutation (at the corresponding optimal screening age) for different VAF thresholds ranging from 2% to 0.01% (x-axis in log scale).

**Acknowledgements:** P.-H.C. and I.P. received grants for the Prism Project, funded by the Agence Nationale de la Recherche under grant ANR-18-IBHU-0002, and for the Appel a Projets Pre-neoplasies 2021 (C21021LS) with financial support from Institut Thematique Multi-Organismes (ITMO) Cancer of Aviesan within the framework of the 2021-2030 Cancer Control Strategy, on funds administered by INSERM. This work was also supported by grants from INCA Plbio2021 and Ligue Nationale Contre le Cancer (équipe labellisée 2019) to I.P. A.T. was a recipient from MENRT grant then from Ligue Nationale Contre le Cancer. We thank D. Madhavan and M.-L. Charpignon for their help in editing the manuscript.

**Author contributions:** G.H., I.P., P.-H.C. conceived the mathematical model, G.H., A.R., M.B., P.-H.C. defined and performed the statistical methods, M.B. and P.-H.C. developed the ABC-SMC Julia package, M.E.K, A.T., I.P., conceived and performed the experiments, C. M. performed targeted NGS analysis, F. G. followed patients and provided clinical data, W.V. advised the study, G.H., I.P., P.-H.C. wrote the manuscript, I.P, P.-H.C. supervised the study, all authors revised the manuscript.

**Conflict-of-interest disclosure:**

The authors declare no competing financial interests.

A.

*1- Modelling*

a) Clonal expansion N(t) over time t

b) Two hypotheses of mutation acquisition:

- Occurrence in fetal life ($T_0 = 0$)
- Acquisition after birth ($T_0 > 0$)

*2- Parameter estimation using ABC-SMC*

a) Repeat until *M* parameter vectors are selected

(i) Propose a parameter vector and simulate the model

(ii) Confront with observations → distance d

(iii) If d < tolerance, then: select the parameter vector
else: go back to (i)

b) Sequentially reduce the tolerance and go back to a) until a minimum tolerance is reached.

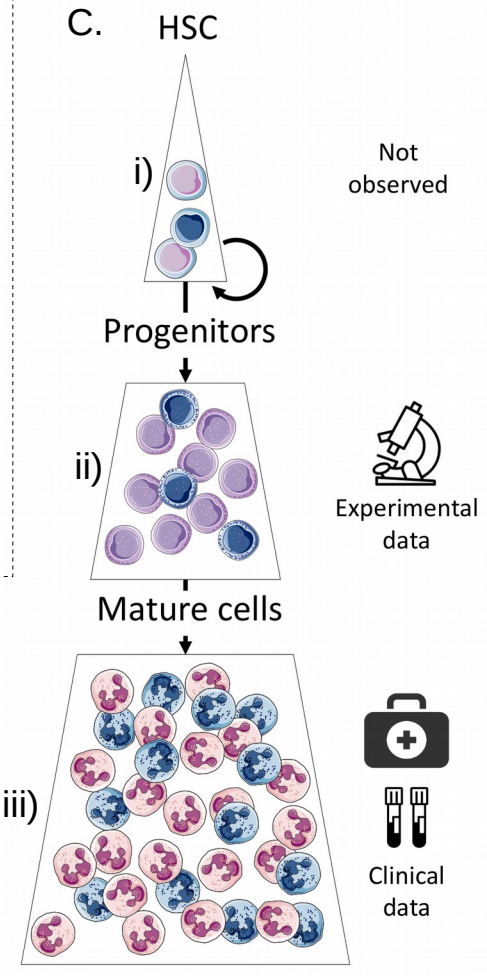
c) Approximate the posterior distribution

*3- Inferring the MPN development*

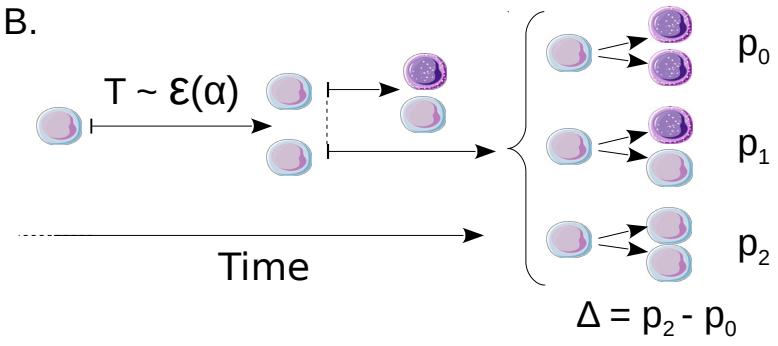
a) Compare *JAK2*$^{V617F}$ to *CALR*$^{m}$

b) Determine early screening strategies

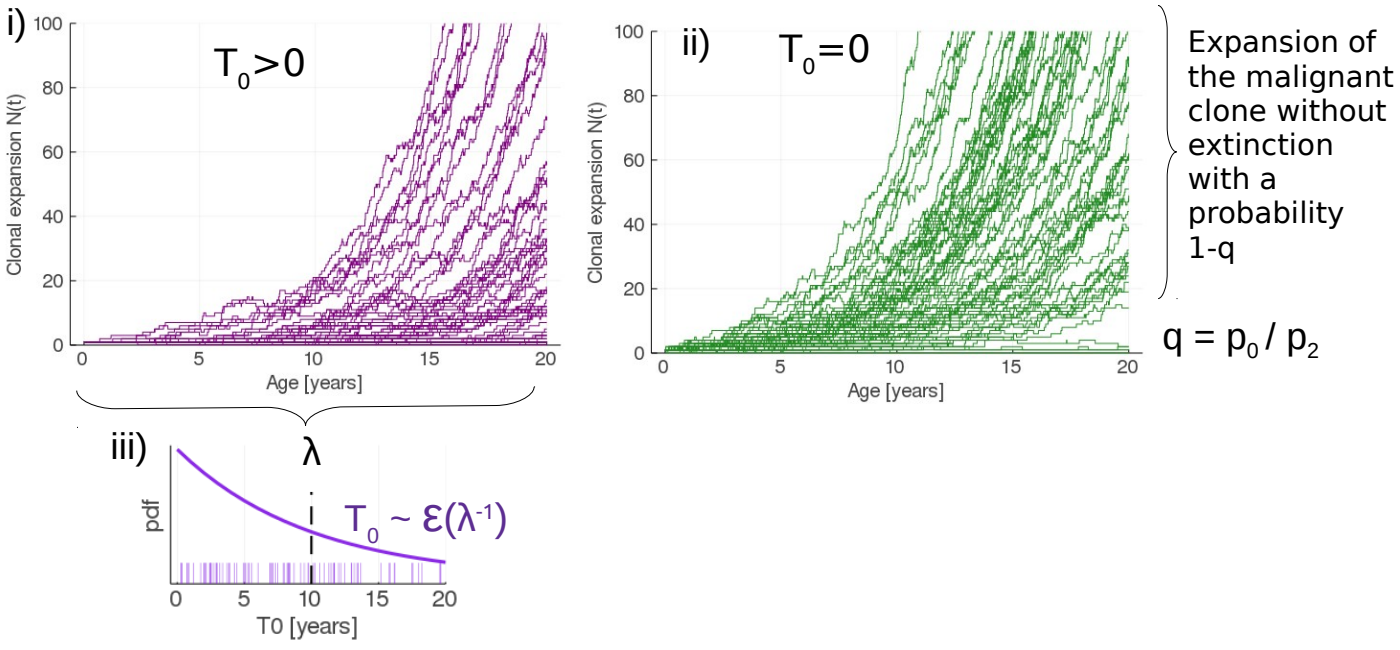
C.

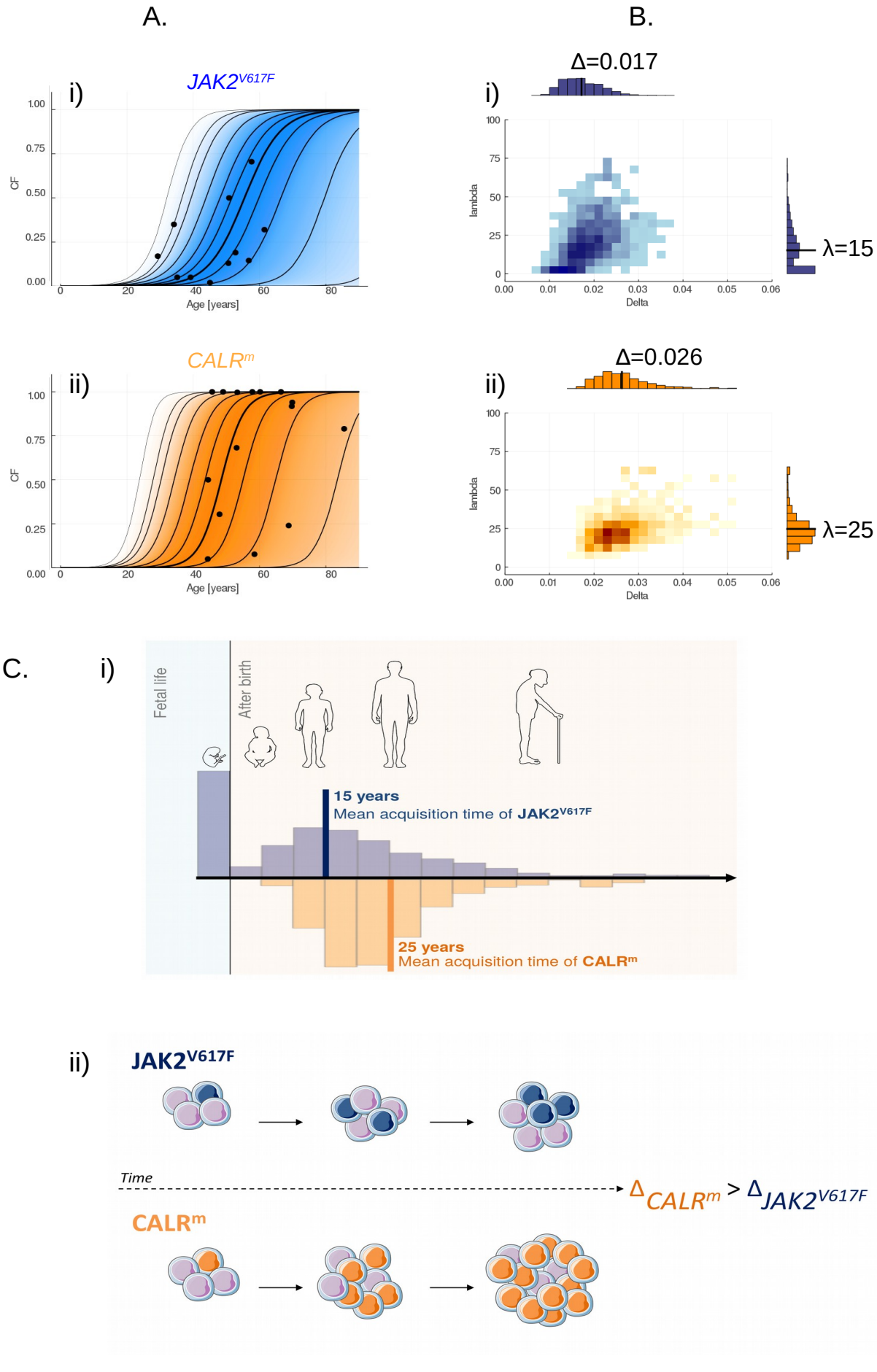


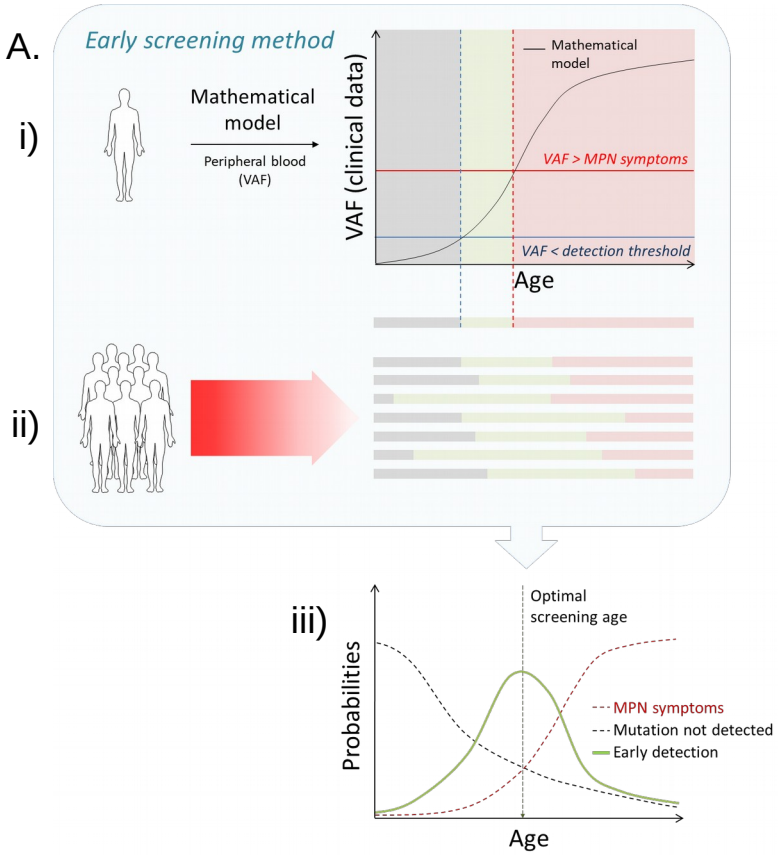

B.

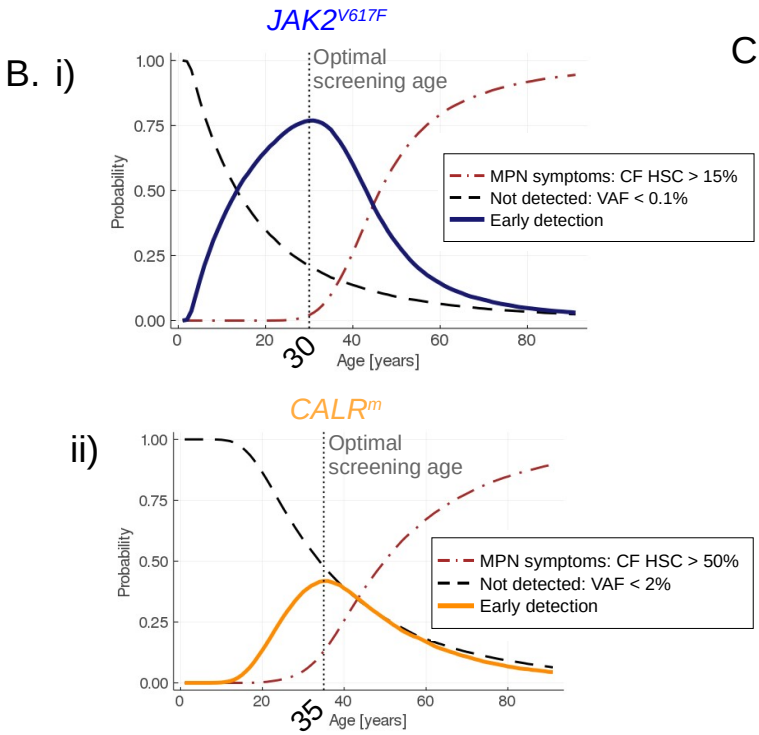


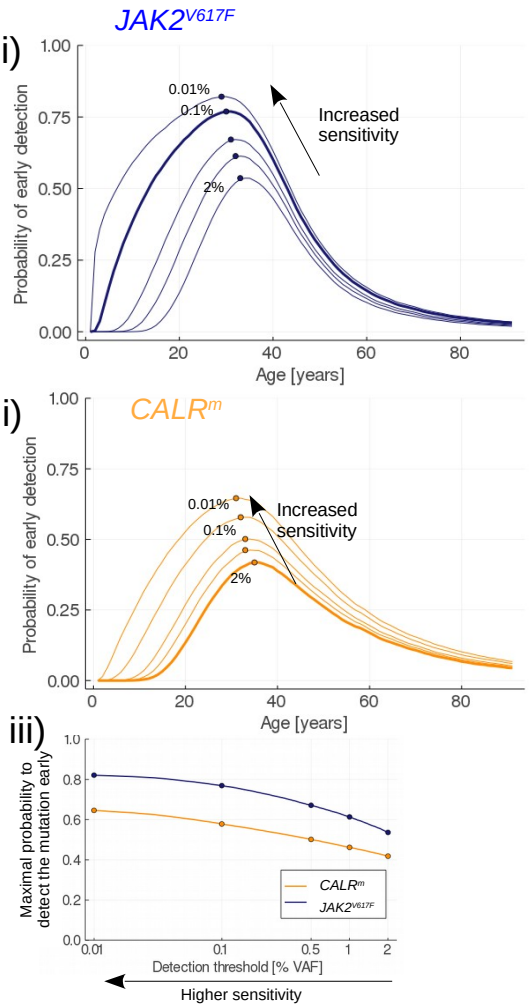

D.




A.
B.
i)
JAK2V617F
CF
Age [years]
Δ=0.017
lambda
Delta
λ=15
ii)
CALRm
Δ=0.026
λ=25
C.
i)
Fetal life
After birth
15 years
Mean acquisition time of JAK2V617F
25 years
Mean acquisition time of CALRm
ii)
JAK2V617F
Time
ΔCALRm > ΔJAK2V617F
CALRm


A.
Early screening method
i)
Mathematical model
Peripheral blood (VAF)
VAF (clinical data)
Mathematical model
VAF > MPN symptoms
VAF < detection threshold
Age
ii)
iii)
Optimal screening age
Probabilities
MPN symptoms
Mutation not detected
Early detection
Age


B. i)
JAK2V617F
Optimal screening age
Probability
MPN symptoms: CF HSC > 15%
Not detected: VAF < 0.1%
Early detection
30
Age [years]
ii)
CALRm
Optimal screening age
Probability
MPN symptoms: CF HSC > 50%
Not detected: VAF < 2%
Early detection
35
Age [years]


C. i)
JAK2V617F
Probability of early detection
0.01%
0.1%
2%
Increased sensitivity
Age [years]
ii)
CALRm
Probability of early detection
0.01%
0.1%
2%
Increased sensitivity
Age [years]
iii)
Maximal probability to detect the mutation early
CALRm
JAK2V617F
Detection threshold [% VAF]
Higher sensitivity