POLITECNICO DI TORINO

Master Degree course in Physics of complex systems

Master Degree Thesis

Stochastic modelling and statistical inference of hematopoietic development in blood cancers

Supervisors
Gurvan Hermange
Paul-Henry Cournede
Andrea Antonio Gamba

Candidate
Margherita Bruno

Academic Year 2023-2024

Abstract

Myeloproliferative neoplasm emerges from somatic mutations in hematopoietic stem cells, in particular it is strongly correlated with mutation of gene JAK2 that gives cells a competitive advantage over wild-type cells. Understanding the dynamics of mutated cell populations and disease progression could enhance diagnostics and treatment. Due to the impracticality of directly observe hematopoietic stem cell behavior in humans, a mathematical model for cell population development is proposed. This model must capture biological processes' inherent variability and stochasticity, with parameters inferred from real data. This internship builds on previous work modeling the development of JAK2-V617F mutated stem cells, using Approximate Bayesian Computation for parameter inference from patient data. The model's extension involves applying it to data from additional patients and further refining the inference algorithm, in particular concerning the definition of distance that approximates the likelihood in a standard Bayesian framework.

Contents

Chapter 1

Introduction

Hematopoietic stem cells (HSC) are unspecialized cells that give origin to all specific blood cells in the human body. During cell division, somatic mutations, i.e., variation of a part of the DNA sequence, can be acquired at random and then transmitted to the next generation of more differentiated cells such as progenitors and then precursors and mature cells. Some of these mutations can give anomalous features to mutated cells or even competitive advantage with respect to normal cells, particularly when they are acquired in the coding part of the gene and they modify the function of encoded proteins [MM17]. Myeloproliferative neoplasm (MPN), a type of blood cancer that results in the overproduction of mature blood cells, arises following the acquisition of somatic mutations in hematopoietic stem cells. Some individuals with MPN present higher number of red blood cells (polycythemia vera [PV]), others more platelets (essential thrombocythemia [ET]), and some fibrosis of the bone marrow (primary myelofibrosis [PMF]) [ELP+05]. The mutually exclusive driver mutations correlated with this disease happen in three distinct genes: JAK2, CALR and MPL, and they give mutated cells a selective advantage over normal cells (wild-type WT), allowing the mutation to be transmitted to descendants and resulting in the growth of the mutated cell population [JCEea13].
In particular mutation JAK2-V617F is found in 70% cases of MPN [JGD+06] and is correlated with increased production of mature blood cells of the myeloid lineage. Inferring the dynamics of the population of mutated cells and the history of disease progression in an individual could give information about the link between the two and could help in the diagnostic and the cure [VK17]. For MPN it is believed that the malignant mutation is acquired by the patients years before arising of symptoms and following diagnosis [WLM+22], leading to a slow development trough years of the population of mutated HSC that could ideally be detected before leading to the disease, technically referred as early screening.

Since it is unfeasible to directly measure human HSC dynamics and determine the time of onset of the driver mutation, a mathematical model for cell population development can be proposed. A mathematical model to describe this phenomenon has to capture the variability and stochasticity typical of biological processes while explaining the complexity of it.

The mathematical model can consider continuous-time dynamics [WLM+22] or discrete ones [Wri31] and multiple approaches relying on stochastic and branching processes or not [Fis23]. The parameters of a given model are usually not directly accessible, so they have to be inferred based on real data, considering that the estimated values should not be dependent on the procedure chosen while still allowing the reconstruction of the behavior shown in the data from the model [HRB+22]. Different approaches are possible, depending on the patient data available and on the assumptions made about the biological process.

The inference procedure can be carried out in a patient-specific way [VEEN+21] or include multiple patients [HRB+22].

This internship first follows the study conducted in [VEEN+21], that aims to model the development of a population of mutated stem cells over time, from birth to last sample time in a patient, that carry the mutation JAK2-V617F. This is done in two patients with essential thrombocythemia, ET1 and ET2, by analyzing the genealogy and lineage history of a group of HSCs and multipotent progenitors (MPP). The novelty of the approach relies on the choice of using whole genome sequencing and constructing a phylogenetic tree to analyze the development of the population of mutated cells, which requires a particular choice of model and inference. A mathematical model that describes it as the result of stochastic processes is proposed and its parameters are inferred using a Bayesian approach based on patient data. Since the computation of the likelihood of the model parameters given the experimental data is unfeasible for such model it is used an Approximate Bayesian Computation (ABC) method, to reach an approximation of the posterior distribution for the model parameters. This method relies on the approximation of the likelihood with a distance function, to be defined, between experimental data and synthetic data produced by a simulation with given parameters. It is particularly used to infer parameters of biological and population models, since in many of these cases the process can be efficiently simulated. In this case dynamical behavior of the mutated cells is studied from the reconstructed lineage history of the mutation and its development from single-cell genomic and transcriptomic measurements to reconstruct the history of the disease across a patient lifetime. We reproduce the results from [VEEN+21] on the patient data mentioned, highlighting the assumptions made and the possible limitations, in particular regarding the choice of distance in the ABC procedure and the limited number of patients studied.

To tackle these limitations, for the second part of the work we extend the method proposed by [VEEN+21] by applying it to data from different patients studied in [WLM+22] and by considering further analysis using some modifications of the ABC algorithm, in particular exploring different definitions of distance used. In addition the use of Sequential Monte Carlo ABC is investigated, to open the possibility to give more precise and faster results.

Chapter 2

Data analysis and model definition

2.1 Reconstructed lineage from data

The data available, taken from [VEEN+21], consist in Somatic Single-Nucleotide Variants detected in single-cell transcriptomic profiling of a given number of HSC and progenitor cells coming from a given patient with MPN, which can be synthetized as a genotype matrix. The group of cells considered was a mixture between HSC and multipotent progenitors, which are their direct descendants, and for each cell whole-genome sequencing was performed. The relevant information in the genotype matrix for each cell is if it carries or not a particular mutation (1 for present and 0 for non present) and, since gene mutations are transmitted during cell division, this could give insights about ancestral relationships between cells [LSOS+18]. Indeed, in the space of all possible mutations, the number of mutations not shared could be a measure of the distance between two cells in the population [CMR+21]. The distance matrix is constructed using as metric the absolute value of the number of not shared mutations between two cells and from that, using a neighbor-joining algorithm, the phylogenetic tree with each cell present in the data is constructed for each patient.
In general, a tree is a structure composed by a set of nodes and a series of links that connect them, called edges, which could be weighted or unweighted. In a phylogenetic tree, each node represents a species or an individual and the edges, or branches, describe genealogical relationships between them; the leaves of the tree are nodes that have only one connection with other nodes and the root of the tree is the common ancestor between all nodes, that do not have any parent. In the considered case the genealogical relationships are given by shared mutations and the leaves of the tree are the cells coming from the experimental data that carry some somatic mutations that distinguish them from the others; a node connecting two cells signals the presence of a mutation shared between the two and so the root of the tree represents the original baseline of mutations common to all cells. The length of branches of this phylogenetic tree describes the distance between two nodes in terms of mutations.
The neighbor-joining (NJ) algorithm [SN87] is a distance-based way to construct a tree, given a $n \times n$ distance matrix between leaves of the tree, being n the number of cells in the experimental data. The algorithm starts with a generic star-like tree where each leaf is connected to a central node x and can be summed as follows.

1. Construct the matrix Q defined as

$$Q_{ij} = (n-2)d(i,j) - \sum_{k=1}^{n} d(i,k) - \sum_{k=1}^{n} d(j,k) \tag{2.1}$$

where $d(i,j)$ is the distance between leaf i and leaf j. To define this quantity is not only considered the distance between the two analyzed nodes but also their distance from all the others, defining a measure of not only how close two nodes are but also how much isolated with respect to the general population they are

2. find the pair (i,j) with $i \neq j$ for which Q_{ij} is the smallest, create a new node u which joins i and j and connect it to the central node x
3. calculate the distance between nodes in the pair and the new node: $d(i,u)$ and $d(j,u)$
4. calculate the distance between all nodes and this new node $d(k,u)\ \forall k \neq i,j$
5. repeat the steps substituting the node u to the pair (i,j)

The distance between nodes in the pair and the new node connecting them is given by

$$d(i,u) = \frac{1}{2}d(i,j) + \frac{1}{2(n-2)}[\sum_{k=1}^{n} d(i,k) - \sum_{k=1}^{n} d(j,k)] \tag{2.2}$$

while the distance between the new node and all other nodes is

$$d(k,u) = \frac{1}{2}[d(i,k) + d(j,k) - d(i,j)] \tag{2.3}$$

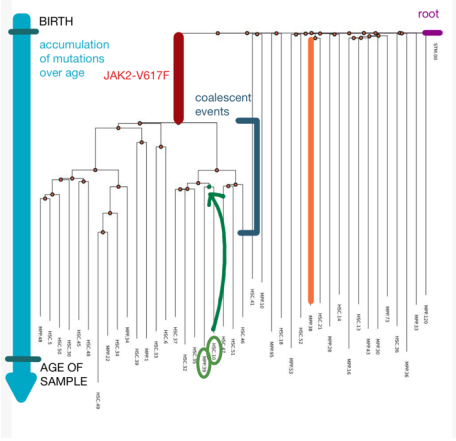
This algorithm can be particularly useful to build phylogenetic trees because in general for a pair of nodes (i,j) the matrix element Q_{ij} is small when the distance between them is small and the distance between them and all other nodes is big. So at each step, the nodes that are connected are the ones that are close to each other but also isolated from the others, meaning that they are closely related in the context of phylogeny.
Also at each step distances are not absolute but they are related to global distances between leaves, maintaining the relative distance between them which in this case is important since it is considered in units of mutations.

Assuming that for these cells there is a constant rate of mutation per year, indirectly, the accumulation of mutations can be seen as a measure of time passed [WLL+12]. Given this and that a shared node signifies the presence of a common ancestor between two cells, it is possible to infer the interval of time of occurrence of a given mutation and to describe its behavior trough time, by exploiting the relation between SNV mutations and time in years.

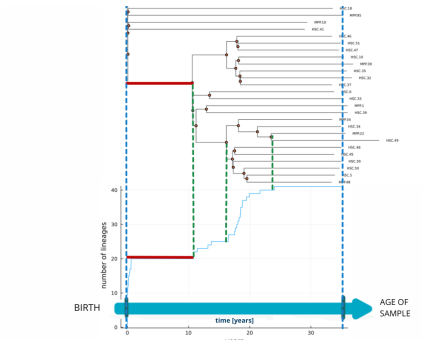
In general there is some variability in the number of mutations acquired by each sampled cell so it is necessary to define an average mutation rate to convert distances on the tree in years. To do so the distance between each leaf of the tree (cell at the time of sampling) and the root, meaning the birth of the patient, is averaged, which corresponds to the average number of mutations that the cells of the patient have accumulated trough life. The average mutation rate is found dividing this quantity by the age of the patient at the time of sampling as

$$l = \frac{\langle number\ of\ mutations \rangle}{age} \tag{2.4}$$

The focus of the work concerns the lineage following the JAK2 mutation, which is known to be responsible for the occurrence of MPN. It is possible to observe a lot of coalescent (lineage division) events in the first years of life, as an example in figure 2.1a and 2.2a are reported the reconstructed trees of the patients analyzed in [VEEN+21].
The lineage that carries the JAK2 mutation differs from the WT lineages for the presence of another growth burst. Since the mutation gives the cells a reproduction advantage with

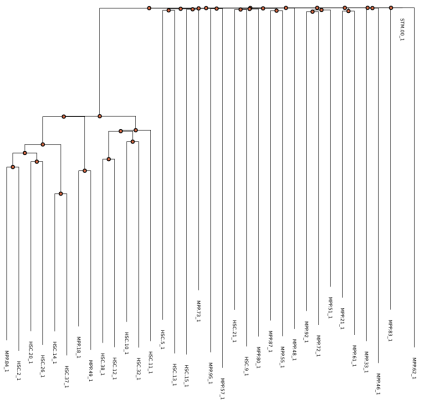


(a) Phylogenetic tree: the root of the tree (in purple) represents the birth of the patient, where zero somatic mutations are assumed, the branches (in orange) are expressed in number of mutations and the leaves are HSC and MPP cells, connected based on genealogical relationships derived by the presence of a common mutation in the data (as an example two leaves are highlighted in green connected to their common ancestor). The JAK2-V617F mutated lineage is highlighted in red and its selective advantage is signaled by a lot of coalescent events (in blue).

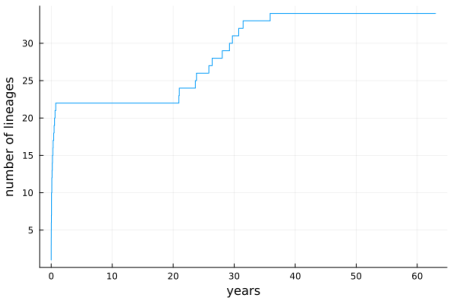


(b) Construction of LTT plot: at each point in time in years is reported the number of present lineages, starting from zero and ending at age of the patient with number of lineages equal to the number of sampled cells. The first coalescent events that give rise to multiple lineages in short time are visible in the first years of life and the division events coming from the JAK2 mutated lineage (in red) are visible after some years (some coalescent events highlighted in green).

Figure 2.1: Schematic of phylogenetic tree and LTT plot based on data from patient ET1 from [VEEN+21].



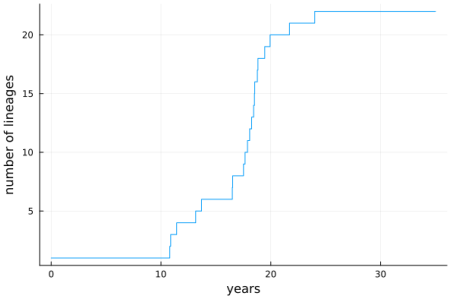
(a) Phylogenetic tree for patient ET2, root of the tree (in purple) represents the time of conception and the leaves of the tree are the sampled cells (in green), the mutated lineage is visible in the centre-bottom of the tree and highlighted in red.



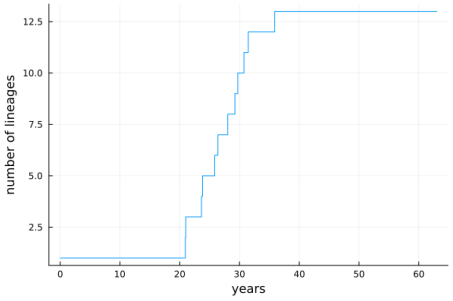
(b) LTT plot for patient ET2: at each point in time in years is reported the number of present lineages,starting from zero and ending at age of the patient with number of lineages equal to the number of sampled cells. The first coalescent events that give rise to multiple lineages in short time are visible in the first years of life and the division events coming from the JAK2 mutated lineage are visible after some years.

Figure 2.2: Phylogenetic tree and LTT plot based on data from patient ET2 from [VEEN+21]

respect to WT cells, there will be a rapid growth of the mutated cell population which is translated in a lot of coalescent events seen in the phylogenetic tree [WLM+22]. Since the phylogenetic tree is also an indirect measure of the time passed between the acquisition of different mutations it is useful to construct a Lineage Trough Time (LTT) plot, that describes the number of cell lineages at different time steps and it gives information about when a mutation that distinguishes between two lineages is acquired but also about the fitness or the selective advantage of each mutation. The LTT plot is obtained in units of years by considering the average mutation rate calculated from the tree and considering the plot up the age of the patient.
From the LTT plots in figure 2.1b and 2.2b is clearly visible the second group of coalescent events coming from the mutated lineage and the different times of acquisition of mutations that differentiates the lineages. The JAK2 mutation is acquired by the patient in a time



(a) LTT plot for mutated lineages for patient ET1



(b) LTT plot for mutated lineages for patient ET2

Figure 2.3: LTT plots considering only mutated lineages for patients ET1 and ET2 from [VEEN+21]. At each point in time is reported the number of lineages that carry the JAK2 mutation, from zero to the number of mutated cells sampled; the driver mutation could have been acquired somewhere between zero and the first coalescent event and how lineages are acquired trought time is related to the fitness advantage given by the mutation.

between the birth and the clonal outburst signaled by a lot of coalescent events, as it is possible to see from the LTT plots constructed considering only the mutated lineages as in figure 2.3. If the fitness of the mutation is high then there will be a lot of coalescent events in short time, signifying a big clonal expansion. So the shape of the LTT plot contains information about when the mutation was acquired and the reproductive advantage of the mutation with respect to WT cells. On the other hand in the LTT plot there is no longer information about the topology of the tree and genealogy relations between lineages of cells.

2.2 Simulations using Wright-Fisher model

The mathematical model to be considered has to describe the development of a population of mutated cells among wild-type non-mutated ones, starting from a single mutated cell. In particular what is studied is the number of mutated stem cells n_t as a function of time t.
The Wright-Fisher model [Wri31] - [Fis23] is a mathematical model of cell renewal described as a time-discrete stochastic process. The model assumes that the number of cells in the population is constant and that each cell is identical and independent, both from all cells in the same generation and from cells at the previous generations. This model can be easily implemented and simulated while controlling precisely the parameters and

the assumptions made, which is useful in the context of approximate inference since a lot of simulations have to be performed.
The parameters considered are: N, the maximum number of mutated stem cells, t' age of appearance of the first mutated cell and s fitness of the mutation. N is the total number of stem cells considered in the population, which are all wild-type for every time $t < t'$ before the mutation appears, and s refers to the proliferative advantage that mutated cells have over WT ones at every new generation. The self-renewal process is considered to be lasting for L generations, then being L the age of the patient in years, while the mutated lineage develops for g generations, given $t' = L - g$ and one generation per year considered. At the beginning of the process, the population of mutated cells has to survive stochastic extinction and then they grow almost following a deterministic law. The model assumes that growth only depends on the fitness and the number of generations and not on the cell division rate, which is considered constant.
At every generation, N cells are generated, then each cell is assigned a parent at random from the previous generation and inherits its genotypic state, continuing the lineage. If no mutation is acquired at each generation N WT cells are generated. At time t' the mutation is acquired and so the number of mutated stem cells is $n_{t'} = 1$. For time $t > t'$ each of the N new cells has a given probability of choosing a mutated cell as its parent and so becoming mutated as well.
The proliferative advantage of the mutated cells is encoded in parameter s by considering the probability of a given cell to be selected as parent. If each WT cell has probability p of being selected as parent then each mutated cell has probability $(1 + s)p$ to be selected. Given the n_{t-1} number of mutated stem cells in the last generation $t - 1$, the probability for a cell in the new generation t to choose a WT type cell as parent is $(N - n_{t-1})p$ while it is $(1 + s)n_{t-1}p$ for mutated cells; given that the sum of these probabilities has to be 1 for each cell it is possible to obtain a value for the parameter p as $p = \frac{1}{N+sn_{t-1}}$.
If one considers that one cell inherits its status, mutated or not, from its parent, at each generation t the probability of a cell to be mutated is $p_t^{mut} = \frac{(1+s)n_{t-1}}{N+sn_{t-1}}$ which depends on the number of mutated cells at the previous generation and is the probability for a cell to choose a mutated cell as its parent.
So given the number of mutated cells at the previous generation, each cell is mutated or not independently and the number of mutated cells at a given generation follows a binomial distribution of parameters N, the total number of cells, and p^{mut} the probability to be mutated for each cell.

$$P_N(n_{mut} = k) = (p^{mut})^k(1 - p^{mut})^{N-k} \tag{2.5}$$

The mean value of the number of mutated cells at each time step given the number of mutated cells at the previous generation can be computed.

$$E[n_t|n_{t-1} = n] = N\frac{(1 + s)n}{N + ns} \tag{2.6}$$

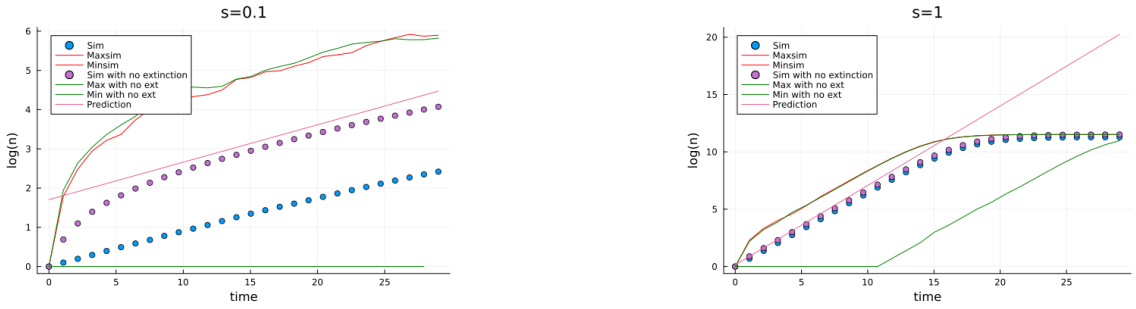
assuming $N >> ns$ and giving the expectation value as a function of the number of mutated cells at the previous generation it is obtained

$$E[n_t|n_{t-1}] = (1 + s)n_{t-1} \tag{2.7}$$

Then taking the average over the number of mutated cells at generation t-1 on both sides gives rise to a recursion $E[n_t] = (1 + s)E[n_{t-1}]$ which can be solved as

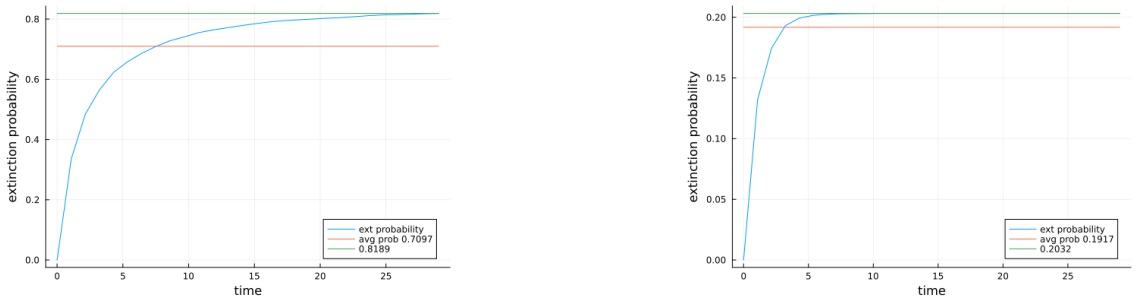
$$E[n_t] = (1 + s)^t \tag{2.8}$$

with $E[n_{t=t'}] = 1$ and it describes an exponential growth.
The process can be simulated for g generations to obtain the number of mutated stem cells at each generation. It is important to notice that at each time step there is a non zero probability of extinction, that is when the number of mutated cells goes to zero. This effect needs to be taken into account since trajectories that exhibit extinction would not be considered when performing simulations for inference.
To do a qualitative analysis of the model for a specific set of parameters, several trajectories are simulated and then averaged, considering both ones that can go extinct and ones that survives stochastic extinction. For different values of the fitness, the mean trajectory exhibits different behaviours. For each mean trajectory 1000 simulations were performed, with or without considering extincted trajectories in the average, for $g = 28$ generations and $N = 10^6$ total cells.



(a) s=0.1 **(b)** s=1

Figure 2.4: Averaged trajectories obtained from 1000 simulations with (purple) and without (blue) excluding extincted trajectories; the deterministic approximation is reported (pink) and the maximum and minimum values for each time step (in red with extincted trajectories and in green without).



(a) s=0.1 **(b)** s=1

Figure 2.5: Extinction probability as a function of time (blue) obtained from 1000 simulations; the the saturation value for the extinction probability at the last time step (green) and the average value of the extinction probability over time (orange) are reported.

In the figures 2.4a and 2.4b are reported respectively simulations for parameter s=0.1 and s=1, considering also the maximum and minimum values of the number of mutated stem cells n_t at each time step and the theoretically predicted trajectories which describe the deterministic behaviour.
For smaller value of the fitness first the mean trajectory takes into account the time of survival of stochastic extinction, then it follows the deterministic predicted curve. The

mean curve for higher values of the fitness shows an exponential increase for the first generations which is consistent with the theoretical calculations in equation (2.8) and then there is a saturation of the number of mutated cells which is when the approximation $N >> ns$ breaks down since then the number of mutated stem cells becomes equal to N. The extinction probability can be calculated empirically from the simulations considering the number of extincted trajectories at each time step. In the first generation is always zero since every simulation starts with fixed $n = 1$ mutated cells, then increases up to a point where it becomes constant in time, with a value depending on the fitness parameter as shown in figure 2.5.

2.3 Reconstructed lineages from simulations

For each simulated trajectory it is obtained the number of mutated cells at each generation without considering any genealogical relationship between them since for every cell at a given generation the probability to be mutated depends only on the number of mutated cells at the past generation. In the patient data there are k cells that carry the mutation so from the simulation k cells are selected at random from the population of mutated cells at the last generation and for them the genealogy is reconstructed.
To do that it is necessary to assign parent-child relationships between cells of different generations. This is done by assigning to each cell one cell of the previous generation at random as parent, which is the assumed process for the simulation. Simulating the process forward without considering genealogical relationships and then reconstruct them backward from the last generation to the first parent cell is equivalent to simulate random genealogies from the start of the mutational process. Then a phylogenetic tree is constructed, defining relationships for all k cells. At this stage a branch of length one connects each node to its parent node so the length of the branches of the tree just signal the passage of generations between coalescent events.
It is important to notice that in the experimental tree the length of each branch represents the number of mutations separating one cell from an ancestor, which is not constant in general, so to incorporate that for the simulated tree the length of each branch is assigned at random from a Poisson distribution, with mean the mutation rate per generation. The Poisson distribution is a discrete distribution and it expresses the probability of a number of independent and successive events happening in a given interval of time, with a given rate λ that it is also the mean value of the distribution. So the probability of n events to happen in the unit of time is

$$P_\lambda(n) = \frac{\lambda^n}{n!} e^{-\lambda}$$

.
The mean mutation rate is specific for each patient and has to be derived empirically from the patient tree as shown in equation (2.4) in section 2.1, expressed in number of mutations per year since it is assumed one generation of cells per year. These two processes allow to construct a tree in the most general way without any other assumption since both the topology of the tree, coming from the random ancestral relationships, and the length of each branch are random. In this way the resulting tree of a simulation with given parameters is not deterministically constructed but the variability typical of biological processes is incorporated.
It is useful to see in figure 2.6 that for various trees with same genealogical relationships but different branch lengths the resulting LTT plot is different, which has to be taken into account in the inference part. In particular the distance between two piecewise constant

functions such as the LTT plots can be computed as the area below the curve defined as the absolute value of the difference at each point and then it is possible to see that it is not zero for the two LTT plots coming from the same tree topology.

It is important to notice how variability resulting from the Wright-Fisher process and

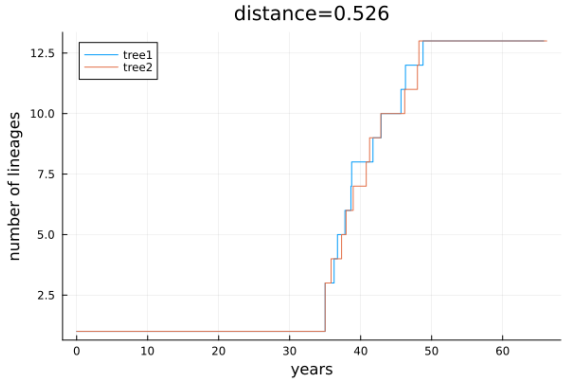
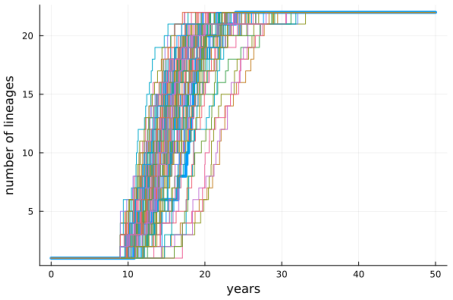


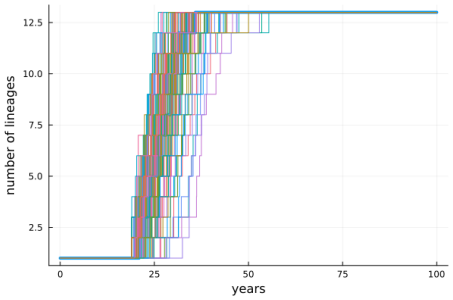
Figure 2.6: LTT plot obtained for trees with same nodes but different branch length, due to the use of two different realizations of the Poisson process for defining branch length,: the distance between them is computed and it is not zero.

from the construction of the tree in the simulation can influence the result in the LTT plot, since the distance from the simulated plot and the data would be the metric to accept or reject a parameter vector.

For the same values of the parameters, there is a big variability between LTT plots. To test that, 100 simulations with the same parameters vectors are performed and it is computed the average distance from the data. This is done for the best estimation of the parameter values given by [VEEN+21]: $g = 26$ and $s = 0.63$ for ET1, and $g = 44$ and $s = 0.44$ for ET2 as shown in figure 2.7.



(a) g=26, s=0.63, LTT plot from patient ET1 is considered: average distance= 2.37



(b) g=44, s=0.44, LTT plot from patient ET2 is considered: average distance= 2.51

Figure 2.7: Comparison between experimental LTT plot for patients ET1 and ET2 (in blue) and 100 LTT plots obtained by simulations with fixed parameters; the parameters chosen were the ones estimated by [VEEN+21] for age of onset and fitness of the mutation for the two patients. The average distance from the data is reported.

Chapter 3

Approximate Bayesian Computation

3.1 Approximate Bayesian Computation

The Approximate Bayesian Computation (ABC) method is an approximation of the classical Bayesian approach on parameter inference. The objective is to estimate the posterior distribution of the parameters of a given model based on observations of the data while incorporating previous knowledge on the parameter values [MMPT03]. In the classical Bayesian approach the posterior distribution of the data is proportional to the likelihood of the parameters, which quantifies how much likely are the data to have been generated by the model with that specific parameters, multiplied by the prior distribution of the parameters, which incorporates any knowledge about the parameters before any observation has been made. Considering a vector of parameters θ and a set of data D:

$$p(\theta|D) \propto \mathcal{L}(D|\theta)p(\theta) \tag{3.1}$$

In many cases the computation of the likelihood can be infeasible so in the ABC method the precise computation of the likelihood is substituted by a simulation of the model with given parameters values, which is then confronted with the data [GF20]. The main assumption is that it is possible to define a parametric model which simulates the mechanism that had generated the data and so for each value of the parameter vector θ it is possible to obtain simulated observations. It is possible to obtain simulated observations of the data based on the mathematical model with a given parameter vector that are compared with the data by defining a distance; the parameter vector is accepted to approximate the posterior distribution if the simulated data are at a distance less then a threshold from the data. In general, this process can be applied to some summary statistics (or function in general) of the data $\eta(D)$.
If the process generating the simulation is stochastic for a summary statistic of the data which is considered regular it is reasonable to approximate its distribution with a Gaussian distribution

$$p(\eta_0|\theta) = \frac{1}{(2\pi)^{p/2}|\Sigma_\theta|^{1/2}} exp(-\frac{1}{2}[(\eta_0 - \eta_{sim}^{(j)})^T \Sigma_\theta^{-1}(\eta_0 - \eta_{sim}^{(j)})]) \tag{3.2}$$

which is centered around the "true" value of the summary statistics for the simulation for those parameters and with a variance that depends on the stochasticity of the model.
This distribution can be used to define a synthetic likelihood function, by considering the

distribution of the simulations with respect to the data.

$$L_s(\theta) = \int K_\epsilon(\eta_0, \eta_{sim})\mathcal{N}(\eta_{sim})d\eta_{sim} \tag{3.3}$$

If a Gaussian distribution is considered also for the distribution of the simulated summary statistic for different parameter values, as a function of the parameter ϵ, which describe the acceptance threshold

$$K_\epsilon(\eta_0, \eta_{sim}) = \frac{1}{(2\pi\epsilon)^{n/2}} exp(-\frac{1}{2\epsilon^2}(\eta_{sim} - \eta_0)^T(\eta_{sim} - \eta_0)) \tag{3.4}$$

then maximizing the likelihood would be the same as minimizing the distance between the summary statistics of the data η_0 and of the simulations η_{sim} which depends on the parameter θ.
In the case studied the summary statistic for the data is the Lineage Trough Time function and the distance is defined as the area below the piecewise-constant curve obtained from the absolute values of the difference between experimental and simulated LTT plots.
Given the prior distribution of the parameter vector, the algorithm would proceed like this [TWS+09]

1. vector of parameters is sampled from the respective prior distributions $\theta = (s, g, L, ...)^t$
2. simulated tree and LTT plot is produced using the Wright-Fisher model for the given parameters
3. distance between the simulated plot and the data plot is calculated
4. if the distance is smaller than ϵ the vector of parameters is kept in the approximation of the posterior distribution, else it is disregarded

This steps are repeated until a sufficient number of parameter vectors are kept in the approximation of the posterior.
The histogram constructed from the retained values of the parameters would be used to approximate the posterior distribution, both for the marginal of each parameter and for the joint probability distribution.

3.2 ABC using average distance

From this section on, some variants to the method used by [VEEN+21] are proposed, to compare the results obtained with different procedures and to investigate some criticality of the model.
As explained in equation (3.2) in section 3.1 due to the stochasticity of the model there is some variability in the results of the simulations for the same parameter vector and this can be mitigated by performing multiple realizations of the simulated summary statistics with the same parameters, to ensure that multiple variants of the simulations are taken into account.
In this case a number N of simulations with the same vector of parameters is considered and the Gaussian distribution of the summary statistic of the data will have parameters $\widehat{\mu_\theta}$ and $\widehat{\Sigma_\theta}$ that are approximated by averages taken over summary statistic of the simulated data [GF20].

$$\widehat{\mu_\theta} = \frac{1}{N}\sum_{j=i}^{N} \eta_{sim}^{(j)} \tag{3.5}$$

$$\widehat{\Sigma_\theta} = \frac{1}{N-1}\sum_{j=1}^{N}(\eta_{sim}^{(j)} - \widehat{\mu_\theta})(\eta_{sim}^{(j)} - \widehat{\mu_\theta})^T \tag{3.6}$$

So that the distribution is

$$p(\eta_0|\theta) = \frac{1}{(2\pi)^{p/2}|\Sigma_\theta|^{1/2}} exp - \frac{1}{2}[(\eta_0 - \widehat{\mu_\theta})^T \widehat{\Sigma_\theta}^{-1}(\eta_0 - \widehat{\mu_\theta})] \tag{3.7}$$

In this case directly averaging the values of the trajectories or the LTT plots would result in the loss of some details of the model so instead it could be useful to define the distance as the average distance between N realization of the summary statistics of the simulated data and of the data. Then the ABC procedure can be summarized as

1. vector of parameters $\theta = (s, g, ...)$ is sampled from the prior distributions
2. N simulated LTTs are produced using the Wright-Fisher model for the given parameters
3. distance $d_j = d(\eta_{sim}^{(j)}, \eta_0)$, with $j = 1, ..., N$ between each simulated plot and the data plot is calculated
4. average between distances from N simulations and the data is calculated $\overline{d} = \frac{1}{N}\sum_{j=1}^{N} d_j$
5. if the average distance $\overline{d}$ is smaller than ϵ the vector of parameters is kept in the approximation of the posterior distribution, else it is disregarded

This would mitigate the effect of the variability in the LTT plots introduced by the random processes leading to the construction of the tree. In general it is difficult to account for this factor since every parameter vector has its own variability when inserted into the simulation.

3.3 Sequential Monte Carlo ABC

One of the biggest drawbacks in classic ABC method is the long computation time due to low acceptance rate of simulations in the approximation of the posterior distribution. This is related also to the acceptance threshold, which still has not to be too high to obtain a meaningful posterior distribution. To obtain a better and more efficient convergence to the posterior, the Sequential Monte Carlo ABC (ABC-SMC) method could be used [GF20] to improve results from the previous sections.
This method arrives gradually at an approximation of the posterior distribution trough a series of steps, where at each one the approximation of the distribution is refined by accepting parameters with decreasing tolerance.
At the end of each step is defined an intermediate distribution given by the parameter vectors θ and their weights $w_t(\theta)$, which is the prior distribution for the first step. Points are sample from the current distribution and then perturbed with a given perturbation Kernel, so that the whole parameter space has some probability to be explored as in a Monte Carlo step [BCMR09]. For the value of the parameter vector sampled the distance between the simulated data and the experimental data is calculated B_t times at step t, then the vector is retained in the next distribution if the distance from the data is less than the threshold distance in at least one realization. Each vector is assigned a weight defined in such a way to be higher for parameters that are closer to the data and higher for parameters that are not likely to be sampled from the current distribution, so that at

each iteration there is no risk to explore only partially the parameter space.
This algorithm can be summarized in this way [TWS+09]:

1. a vector of parameters $\theta^{**} = (s^{**}, g^{**}, ...)$ is sampled directly from the prior distribution if t=0 or $\theta^{*} = (s^{*}, g^{*}, ...)$ from the previous population with given weights w_{t-1}.

2. the parameters are perturbed with a given perturbation kernel $K_t(\theta|\theta^*)$ to obtain a new vector θ^{**}, resampled if the parameters are outside of the support of the prior distribution

3. the distance between the simulated LTT plot and the data one is calculated B_t times for given parameters θ^{**}, to calculate the number of times the distance is smaller than the threshold $b(\theta^{**})_t$; if $b(\theta^{**})_t = 0$ sample another vector of parameters

4. the vector of parameters is added to the current probability distribution and calculate the weight $w_t(\theta^{**})$

5. repeat until N particles are retrieved in the current distribution

6. normalize the weights for the current distribution and repeat for T steps

The quantity $b(\theta^{**})_t$ encodes the number of times the distance between the simulated LTT plot and the data one is smaller than the threshold distance at that step ϵ_t.
The weight of one vector of parameter are calculated in this way:

$$w_t^{(i)} = \begin{cases} b(\theta^{**})_t, & if\ t = 0 \\ \frac{p(\theta^{**})b(\theta^{**})_t}{\sum_{j=1}^{N} w_{t-1}^{(j)} K_t(\theta^{**}|\theta^{*})}, & if\ t > 0 \end{cases} \tag{3.8}$$

where $K_t(\theta|\theta^*)$ defines the perturbation kernel. The perturbation kernel allows to explore with a given probability the area around the parameters, in particular if for a given perturbation $\theta^* \longrightarrow \theta^{**}$ the kernel is small, so the point θ^{**} is not likely to be sampled, the associated weight would be big, allowing to sample from the area around θ^{**} at the next step. This is one big advantage of this method compared to the classic ABC in the sense that regions in parameter space close to the already accepted points would be explored and this allows to refine the posterior distribution and to have a lower rejection rate.
It is also necessary to define a scheme for decreasing the threshold distance for which parameters are accepted in the next distribution. One can calculate the distance between the B_t simulated LTT plots and the data and take the minimum distance as a measure of how far the sampled parameter vector is from the "true" one. Collecting this information for all points present in a given intermediate distribution gives an histogram of the distances and the value corresponding to a given quantile of that can be taken as the next threshold value. Using the minimum distance between many simulations as the representative distance between a parameter vector and the data ensure that the sequence of thresholds would be decreasing and so at the next iteration there will be a more refine approximation of the posterior.
The number of simulations B_t to be performed for a given vector of parameters at step t can also be increased at increasing number of steps using a precise scheme, to have the probability to accept more points in the intermediate distribution. Indeed it is important to notice that from the definition of the weight, all sampled parameter vectors for which the simulated LTT plot is at distance less than the threshold for at least one time out of B_t are retained in the intermediate distribution, values that would have been discarded

in the classic ABC but also in the averaged ABC, where it is the average distance across simulations that has to be less than the threshold. Given the behaviour described in section 2.3 this is important to take into account the variability across simulations in the inference procedure.

Chapter 4

Results

4.1 ABC on patients ET1 and ET2

The method presented in section 2.1 is first applied to two patients from [VEEN+21]. ET1 is a 35 year-old patient with essential thrombocythemia and from which are sampled 22 mutated cells among 41 total and ET2 is 63 year-old patient again with essential thrombocythemia with a sample of 13 mutated cells among a total of 34. Both patients are untreated at the time of sample. From the reconstructed phylogenetic trees reported in figures 2.1a and 2.2a and the Lineage Trough Time plots in figures 2.1b and 2.2b the mutated lineage is visible from the clonal burst towards the end of the tree.
The mutation rate for both patients is estimated from the reconstructed tree as in equation (2.4) leading to

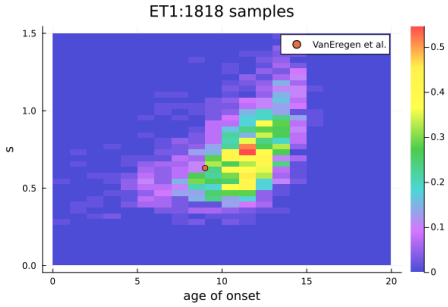
$$l_{ET1} = 20.67655 \tag{4.1}$$

$$l_{ET2} = 19.08868 \tag{4.2}$$

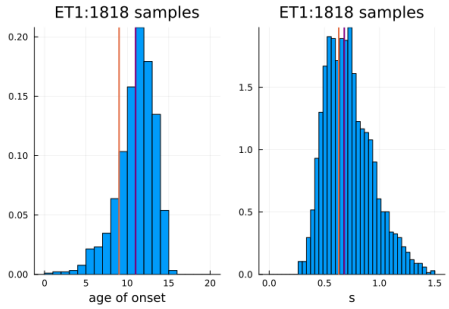
which are consistent with the value $l = 19 \pm 1$ found in [VEEN+21]. In the first analysis the focus will be on the parameters s, regarding the fitness of the mutation, and $t' = L - g$, time of onset of the mutation causing the disease. The other parameters are kept fixed $N = 10^5$, which is an estimation on the total number of stem cells contributing to hematopoiesis [LSOS+18], L the age of the patient, unlike in [VEEN+21], where they perform inference also on these parameters. The value of s is chosen from a uniform distribution between 0 and 2 while g is a natural number chosen from a uniform distribution between 3 and L.
From inspecting the distribution of distances sampling parameters from the prior distribution and the simulated LTT plots obtained as visible in section A, the cut-off distance for patients ET1 and ET2 is chosen as $\epsilon = 1$. With this value and a sufficient number of points an approximation of the posterior distribution for the parameters is obtained and it is consistent with what found in [VEEN+21] for both patients, as it is shown in figure 4.1 and 4.2. For patient ET1 is found that the median values for the fitness is $s = 0.69(0.42 - 1.14)$ and for the age of onset is $t' = 11(6 - 14)$, considering the 90% credibility interval determined by the histogram quantiles.
For patient ET2 the median values are $s = 0.44(0.25 - 0.75)$ and $t' = 20(12 - 22)$.

4.2 ABC on patients PD7271 and PD5163

This procedure, from the mathematical model to the definition of the inference, up until this point has been done following and adapting what has been done in [VEEN+21]. It

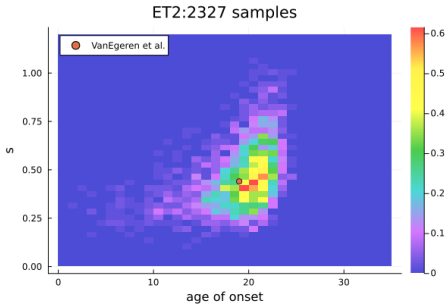


(a) Approximation of joint posterior distribution of parameters s and t' with $\epsilon = 1$. Orange point reports values obtained by [VEEN+21].

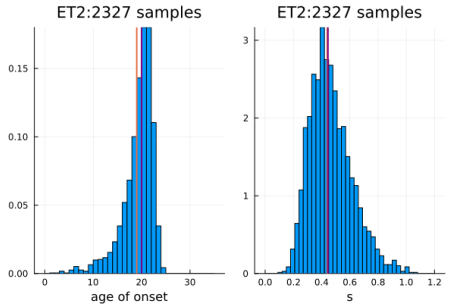


(b) Approximation of marginal posterior distribution for parameters s and t'. Orange line reports the value obtained by [VEEN+21] and purple line reports the median value of the approximate posterior distribution.

Figure 4.1: Results of ABC inference on patient ET1



(a) Approximation of joint posterior distribution for parameters s and t' with $\epsilon = 1$. Orange point reports value obtained by [VEEN+21]



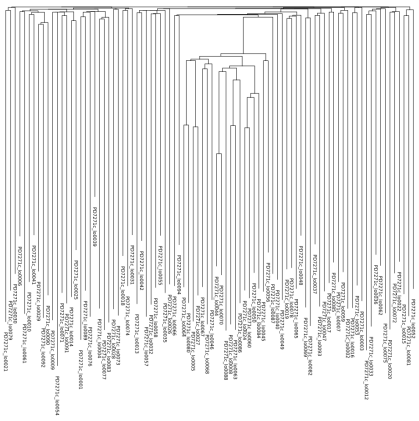
(b) Approximation of marginal posterior distribution. Orange line reports the value obtained by [VEEN+21] and purple line reports the median value of the approximate posterior distribution.

Figure 4.2: Results of ABC inference for patient ET2

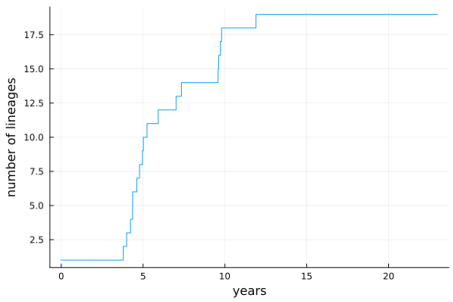
is interesting to apply the same procedure to other similar data to find limitations and strengths. In [WLM+22] the data available are of the same type and the procedure followed using ABC is the same, while the model used is different from the Wright-Fisher and the summary statistics of the data incorporate the LTT plots and the clonal fraction of mutated cells at age of sampling. The patients selected for the testing of the same methodology, described previously in sections 2 and 3 and used on patients ET1 and ET2, were two among the study that carry the disease-initiating mutation JAK2-V617F as in the case of [VEEN+21], that had data collected at the same point in time which is referred to as L, age of the patient, but in some cases treated, which was not the case for the previous patients.

It has to be taken into account that the model used in [WLM+22] considers continuous time while the Wright-Fisher one studies the age of onset on a discrete scale, considering only years, which is a limitation of the Wright-Fisher model. In addition for the construction of the phylogenetic tree in the original paper it is not assumed constant mutation rate across life but instead branch-specific and clade-specific mutation rates are considered, which can modify the association between number of mutations and time in years. Also in [WLM+22] it is suggested that the onset of the mutation JAK2 could have happen before birth, so it is consider that the common ancestor between all cells is dated at time of

conception; this case where the mutation is acquired before birth is not taken into account in the procedure used in [VEEN+21] since the maximum number of generations for the development of the mutated cells population is the age of the patient. To take into account this modification for the new patients 20 mutations, which is the average number of somatic mutations acquired between conception and birth found in [WLM+22], are added to the simulated trees, so that they can be exactly compared with the experimental ones. Patient PD7271 is a 23 year-old patient that presented asymptomatic isolated thrombocytosis and was cured with aspirin, which do not modify the number of somatic mutations present. In this case the genomic matrix presented some missing information about the presence or not of a given mutation in the sampled cell but since it only represents the 0.2% of the total, these mutations at this stage could be considered as not present in the cell. From the phylogenetic tree in figure 4.3 it is possible to notice some differences with



(a) Phylogenetic tree for the PD7271 patient; root of the tree represents the time of conception and the leaves of the tree are the sampled cells, the mutated lineage is visible in the centre-bottom of the tree.



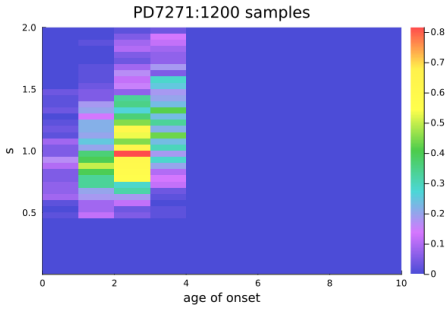
(b) LTT plot for PD7271 patient considering only mutated lineages

Figure 4.3: PD7271, 23 year-old patient from [WLM+22]

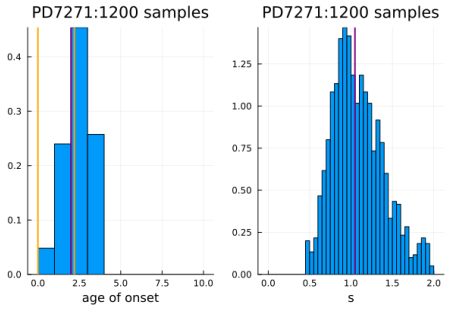
respect to trees from before. The number of cells sampled is higher and as a consequence the number of coalescent events near the root of the tree is higher since with more cells it is more probable that they share a common ancestor later in time and that is visible in the phylogenetic tree. The mutated lineage is clearly visible since apart from this clade the cells shared no mutations close to the end of the tree.
The value of the average mutation rate found for this patient as presented in equation (2.4) in section 3.1 is

$$l_{PD7271} = 20.438 \tag{4.3}$$

which is difficult to be compared with what obtained in [WLM+22], since they consider an average population mutation rate based on multiple patients. Still the value is close to what found for the previous patients in section 4.1 and in [VEEN+21].
The results of the inference that can be compared with the results from [WLM+22] are mainly related to the age of onset of the mutation, since the fitness of the mutation have a different interpretation within the model considered. The first difference noted with respect to the data considered before is that a smaller distance threshold ϵ have to be used to obtain a better approximation of the posterior distribution for the parameters. A distance threshold of $\epsilon = 0.5$ is used in this case. The values obtained for the approximation



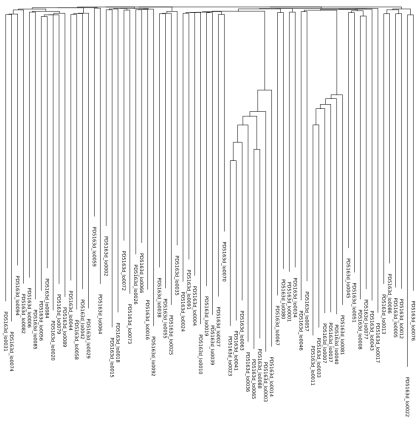
(a) Approximation of joint posterior distribution for parameters s and t' with $\epsilon = 0.5$.



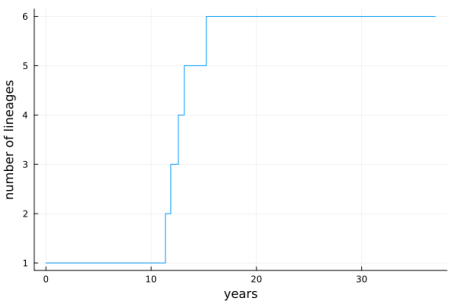
(b) Approximation of marginal posterior distribution for parameters s and t'. In orange is reported the credibility interval for the age of onset from [WLM+22] and in purple the median values of the distributions.

Figure 4.4: Results of ABc inference for patient PD7271

of the posterior distribution are different from the ones reported in the paper as shown in figure 4.4. The median values obtained from the approximation of the posterior are $s = 1.047(0.62 - 1.64)$ and $t' = 2(0 - 3)$.
Patient PD5163 is 38 year-old patient with splanchnic vein thrombosis, a raised red cell mass and received interferon-α from age 31, age of diagnosis, which may influence the number of mutations considered for constructing the phylogenetic tree. Also in this case there are some missing values in the genomic matrix, that represents the 0.25% of the information and so the corresponding mutations are considered as not present in the corresponding gene.
From the figure 4.5 it is possible to notice in the tree two different clades that carry a



(a) Phylogenetic tree for patient PD5163; the root of the tree is the common ancestor between all cells, date at time of conception, while the leaves of the tree are the sample cells. There are two lineages that can be distinguished towards the bottom of the tree, the one considered only refers to JAK2 and it is on the right.



(b) LTT plot for patient PD5163 considering only mutated lineages

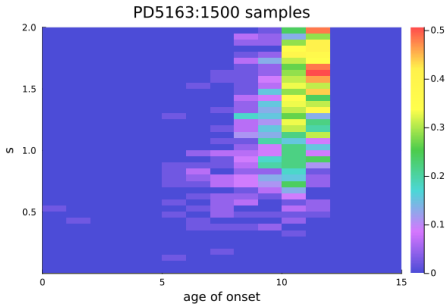
Figure 4.5: PD5163, 38 year-old patient from [WLM+22]

somatic mutation that causes coalescent events in the tree. This work is focused only on

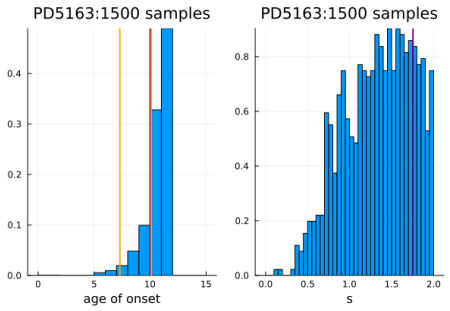
the JAK2 mutated lineage, that concerns six cells and has a lineage history explained in the LTT plot in figure 4.5. It is important to notice that the number of mutated cells considered in this case is lower than the previous ones and this could influence the quality of the inference; still as explored in section 4.7 it would not change drastically the results. The value of the average mutation rate is found to be

$$l_{PD5163} = 19.1107. \tag{4.4}$$

The ABC procedure was performed on the patient with a distance threshold of $\epsilon = 0.4$. An approximation of the posterior distribution is obtained as shown in figure 4.6 and the



(a) Approximation of joint posterior distribution of parameters s and t' with $\epsilon = 0.4$



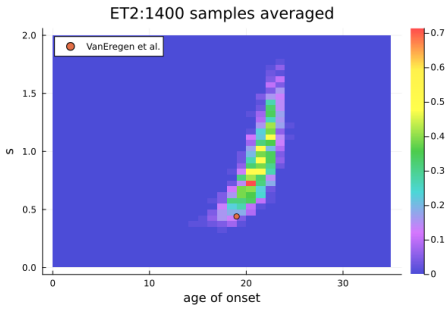
(b) Approximation of marginal posterior distribution for parameters s and t'. The orange lines are the credibility interval from [WLM+22] and in purple are presented the median values of the distributions.

Figure 4.6: Results for ABC inference for patient PD5163

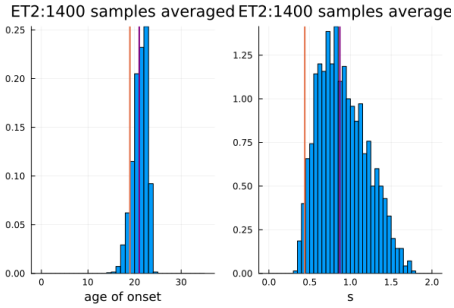
median values are $s = 1.75(0.7 - 2.8)$ and $t' = 10(9 - 12)$.
The results of the inference fall into the credibility interval of [WLM+22] for the age of onset but on the boundary of it and this could be due to the fact that the distance threshold for accepting parameters in the posterior distribution is too high. In particular the median results for the fitness have higher values compared to the ones obtained for patients ET1 and ET2 and its posterior distribution is not as well defined.

4.3 ABC with average distance on patients ET2 and PD7271

To increase the precision of the inference procedure and to take into account the variability of the simulations from the model as explained in section 2.3 it could be useful to implement a version of the ABC with a different definition of the distance as explained in 3.2. In general it is difficult to assert the value of the distance ϵ to be considered to obtain an accurate approximation of the posterior distribution for different data and this could also be a way to mitigate the effect of the stochastic processes present in the simulation and in the generation of the LTT plot.
For patient ET2 10 trees are produced for given values of the parameters and then it is taken the average distance between the respective LTT plots and the data.
Since the summary statistics used now is not anymore the same, given the fact that now it considered the average distance between multiple realization and the data, also the distance threshold has to be re-calibrated. It this case is chosen a distance $\epsilon = 2$.



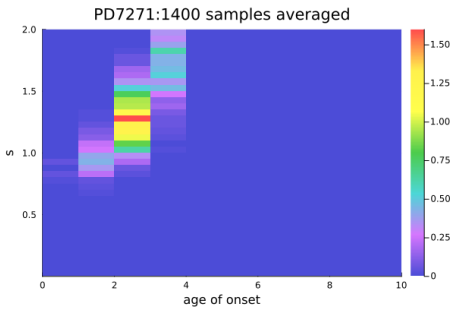
(a) Approximation of joint posterior distribution for parameters s and t' with threshold value $\epsilon = 2$. Orange point reports value obtained by [VEEN+21], which is not well included in the distribution



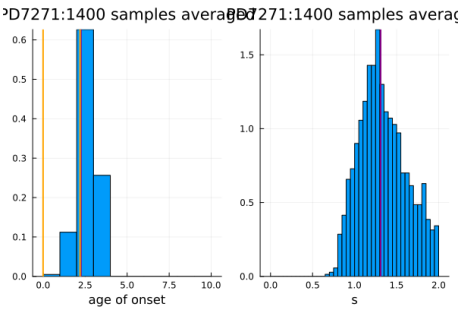
(b) Approximation of marginal posterior distribution for parameters s and t' with threshold value $\epsilon = 2$. Orange line reports the value obtained by [VEEN+21] and purple line reports the median value of the obtained posterior distribution.

Figure 4.7: Results of ABC inference with average distance for patient ET2

The approximation of posterior distribution is shown in figure 4.7 and the median values are $s = 0.87(0.47 - 1.44)$ and $t' = 21(18 - 23)$. What is found in this case is not consistent with the previous result, in particular for values of the fitness. This could be due to the fact that a bigger distance threshold is chosen in the inference with respect to the previous case, but it is useful to notice how changing the definition of distance can change the results in an unexpected way.
For patient PD7271 using the average distance could allow to reach a better convergence of the posterior distribution for parameter s, which could give results closer to the one obtained from [WLM+22]. For each vector of parameters 10 trees were simulated and then their average distance from the data compared with the threshold distance of 1.
The results are shown in figure 4.8 and the median values are $s = 1.3(0.91 - 1.85)$ and $t' = 2.25(1.25 - 3.25)$.



(a) Approximation of joint posterior distribution for parameter s and t' with threshold value $\epsilon = 1$



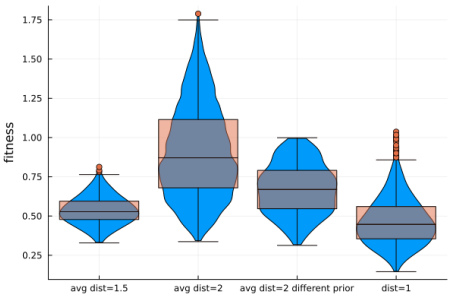
(b) Approximation of marginal posterior distribution for parameters s and t' with threshold value $\epsilon = 1$. Orange lines report confidence interval for age of onset from [WLM+22] and purple lines are the median values for the distribution.

Figure 4.8: Results of ABC inference with average distance for patient PD7271

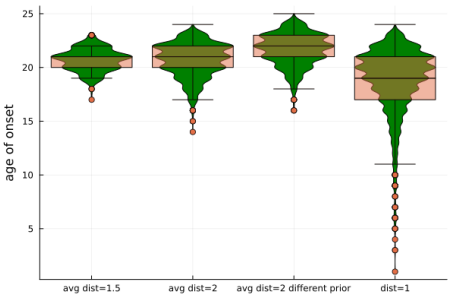
4.4 Bias when using average distance

The results from section 3.2 are different from the ones obtained in section 4.1, which is explained by the fact that a different definition for the distance is used. Since distance defines an approximation of the likelihood of the parameters with respect to the observed data, different definitions of distance would lead to approximations that could vary in principle.
This section investigates the possibility that the use of the average distance introduces a systematic bias in the model and aims to assert the quality of these two definitions of distance.

Firstly, different procedures for the inference of parameters of patient ET2 are consid-



(a) Approximation of the marginal posterior distribution of the parameter s for patient ET2 for different inference procedure



(b) Approximation of the marginal posterior distribution of the age of onset for patient ET2 for different inference procedure

Figure 4.9: Comparison between different approximation of posterior distributions obtained using different procedures. The first and the third quantile of the distribution are reported as the orange boxes, the median is reported and the outliers are reported as orange points.

ered: taking the average distance between multiple simulations and the experimental data with threshold distance $\epsilon = 1.5$ and $\epsilon = 2$, taking the average distance with threshold $\epsilon = 2$ but using as prior for s a uniform distribution between 0 and 1 and the classical ABC considering the distance between one realization of the simulation and the data with distance threshold $\epsilon = 1$ as done in section 4.1. All the distributions found using the average distance are clearly biased towards higher values both for the fitness and for the age of onset as visible in figure 4.9.

It could be useful to inspect the distribution of the distances in the three cases as shown

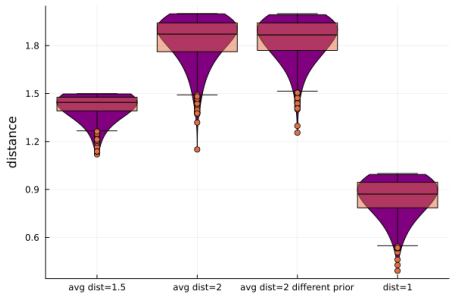


Figure 4.10: Distribution of the distances for the parameter vectors accepted in the approximation of the posterior distribution for different for different realization of the inference

in figure 4.10. It is important to notice that when taking the average the distance between the parameter vector for which are produced multiple simulations and the experimental data never goes below $\overline{d} = 1$ which is actually the threshold value used in the classical ABC, while in that case there are some simulations for which the distance from the data goes towards zero.
To determine the quality of the inference and of the approximations, it is useful to perform ABC with different distances on synthetic data, where the true parameters to be estimated are known. It has to be taken into account that, as explained in section 2.3, the intrinsic stochasticity of the model and tree construction leads to different LTT plots for the same parameter values considered. The result of the inference on synthetic data could then be heavily influenced by the specific realization of the LTT plot used as data. To investigate that, we generated 20 synthetic datasets for the same parameters and we performed ABC considering the two definitions of distance on them. Using this procedure for multiple combinations of parameters could also lead to look for relations between the parameter values and the quality of the inference.

To test the quality of the result, we define a score to be assigned to the inference performed with a specific vector of parameters and a specific definition of the distance. Since the results on patient data are also given in terms of the 90% credibility interval, one possible score could be defined in terms of how many times, on average, the true value of the parameters falls into the credibility interval of the approximation of the posterior distribution.

$$score = \langle \mathbf{I}[\theta^* \in p(\theta)] \rangle \tag{4.5}$$

Where the average is performed over the results obtained from different datasets with the same parameter values.
The maximum value for this score is 1 and it indicates that, for all realizations of the data coming from the same parameter vector, the true value is included in their credibility interval.
The values used for the simulation are $s = [0.1, 0.4, 0.6, 1, 1.5]$ and g=$[30, 33, 35, 36, 37]$ for $L = 40$, $l = 19$ and $k = 15$ and it is performed ABC with the threshold set for every dataset as the 0.01 quantile of the distribution of the distances over 10^5 points sampled from the prior distribution so that for every dataset the number of points considered in the approximation of the posterior distribution is the same.
The results of the score defined in equation (4.5) are reported in tables 4.1 and 4.2 for inference using only one LTT plot to compute the distance and in tables 4.3 and 4.4 for inference when average distance across LTT plots is used.

	30	33	35	36	37
0.1	0.85	0.9	0.8	0.95	1.0
0.4	0.8	0.9	0.85	0.85	0.7
0.6	0.85	0.85	0.9	0.9	0.75
1	1.0	1.0	1.0	1.0	0.95
1.5	1.0	1.0	1.0	1.0	1.0

Table 4.1: Quality score for parameter s

	30	33	35	36	37
0.1	0.85	0.85	0.85	0.8	0.75
0.4	0.85	0.85	0.75	0.85	0.95
0.6	0.9	0.9	1.0	0.9	0.85
1	0.9	0.9	0.9	0.9	0.7
1.5	0.7	0.75	0.9	0.6	0.7

Table 4.2: Quality score for parameter g

It is easy to see that the highest scores for both parameters are obtained when using as distance the area between LTT plots between one simulation and the data. In particular, there are some good results where the value of the score is close or equal to 1, using both definitions of distances but in particular for the fitness parameter s.

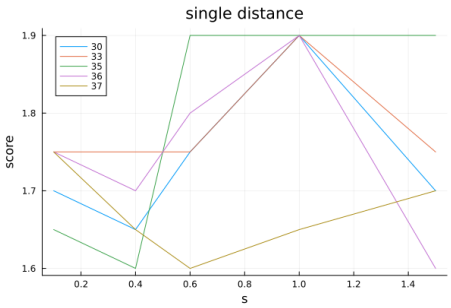
	30	33	35	36	37
0.1	0.55	0.6	0.5	0.7	0.6
0.4	0.3	0.2	0.4	0.3	0.25
0.6	0.2	0.3	0.15	0.4	0.4
1	0.45	0.5	0.5	0.65	0.4
1.5	1.0	1.0	1.0	1.0	1.0

Table 4.3: Quality score for parameter s using average distance

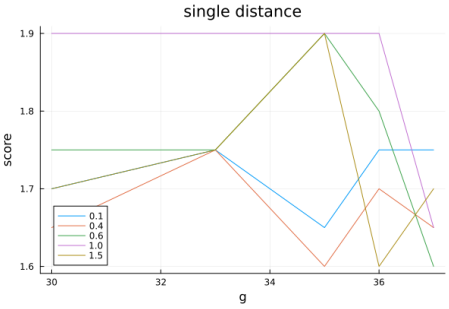
	30	33	35	36	37
0.1	0.65	0.5	0.35	0.65	0.35
0.4	0.5	0.4	0.2	0.4	0.45
0.6	0.35	0.15	0.3	0.35	0.3
1	0.35	0.2	0.25	0.35	0.2
1.5	0.2	0.25	0.4	0.2	0.45

Table 4.4: Quality score for parameter g using average distance

It could be interesting to explore the behavior of the inference depending on the parameter values. Indeed it is possible to sum the score for the same parameter values obtained for s and g, keeping in mind that the maximum value in this case is 2, to have a measure of the global quality of the inference.
The combined score for different parameters is reported in figures 4.11 and 4.12. It is



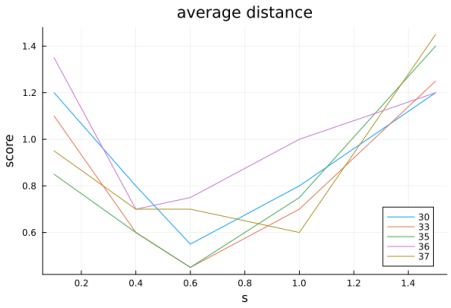
(a) Values of the combined score as a function of the value of the parameter s for different values of g. For almost all values of g the lowest score is found at $s = 0.4$ while there are some good results for $s = 1.0$



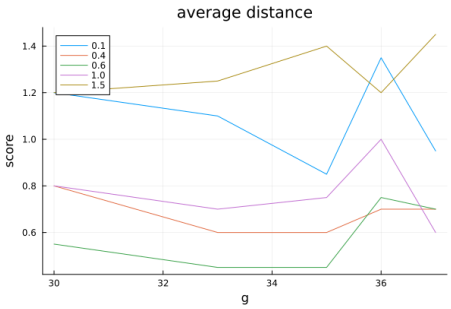
(b) Values of the combined score as a function of the value of the parameter g for different values of s. There is not a particular behaviour as a function of parameters but indeed for $s = 1.0$ the results are good for almost all values of g.

Figure 4.11: Combined score as a function of parameters when performing ABC using single distance

interesting to notice that, in general, across parameter values, the results of the inference when using a single distance can be really different also in terms of quality, while still, both for s and g, have a combined score higher than 1.6.
The combined score when using average distance is lower, meaning that in general the use of his distance could not give a good approximation for the likelihood of the parameter given the data. It is interesting to notice also that the behaviour of the combined score as a function of one parameter is similar across different values of the other parameter, suggesting a common bias when using the average distance.
As an example, the resulting 20 approximations of the posterior distributions are reported in figure 4.13 for parameters $s = 1$ and $g = 35$.
It is possible to notice that, in general, the approximations of the posterior distributions for the same data set with different definitions of distance are coherent, leaning towards values higher or lower than the true value in the same way, and that not for all realization the true value is close to the median value of the resulting distribution. This is due to the variability between different realizations of the LTT plot used for the inference.
To improve the results, all distributions can be combined into one, which takes into account this variability, and indeed, it is observed that the median of this distribution is

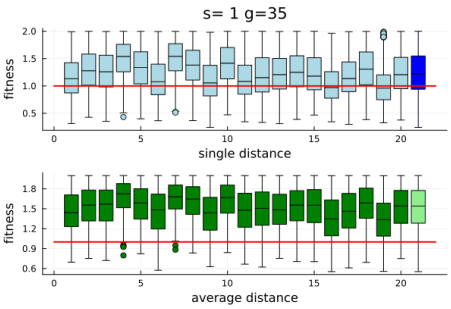


(a) Values of the combined score as a function of the value of the parameter s for different values of g when using average distance. All curves show a similar behavior, meaning that the inference procedure with this definition of the distance has worst performance with $s = 0.4$ and $s = 0.6$.

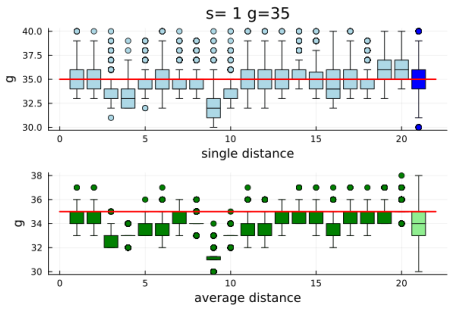


(b) Values of the combined score as a function of the value of the parameter g for different values of s when using average distance. Also in this case the curves have a similar behavior and as expected the lowest score is obtained with $s = 0.6$ across values of parameter g.

Figure 4.12: Combined score as a function of parameters when performing ABC using average distance



(a) Boxplots reporting the approximation of the posterior distribution for the parameter s for 20 datasets obtained from the same parameter values $s = 1$ and $g = 35$ with single (in light blue) and average (in dark green) distance considered. In dark blue and light green are reported the distributions obtained combining all the previous ones and in red is reported the true value that generated the data.



(b) Boxplots reporting the approximation of the posterior distribution for the parameter g for 20 datasets obtained from the same parameter values $s = 1$ and $g = 35$ with single (in light blue) and average (in dark green) distance considered. In dark blue and light green are reported the distributions obtained combining all the previous ones and in red is reported the true value that generated the data.

Figure 4.13: Approximations of the posterior distribution for parameter s and g, the boxplot reports value between the fist and the third quantile and the median. On the same column the inference was performed on the same dataset, with single distance (top) and average distance (bottom). For not all realization of the dataset the true value is close to the median value of the distribution but the result is improved when combining the results into one distribution (dark blue and light green distributions).

closer to the true data. This suggests that a population framework which allows the combination of results coming from multiple realizations of the same process can give a more robust inference.
It is also proven that the variability of the simulations and when constructing the resulting LTT plot influence profoundly the inference procedure, leading to various results.

4.5 ABC-SMC on patient ET2

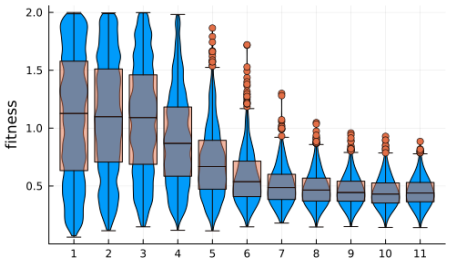
Given that the choice of the threshold distance for the two precedent methods was arbitrary and that this choice influences the results both in terms of convergence to the posterior distribution and in terms of computation time it could be useful to implement the ABC-SMC method on the same data, to analyse also how the distance threshold scales with number of steps and time and to study the convergence to the posterior distribution.
The first testing of the ABC-SMC algorithm is performed using a uniform perturbation kernel both for parameter s and g as

$$K(s|s^*) = Uniform(s^* - \sigma_s, s^* + \sigma_s) \tag{4.6}$$

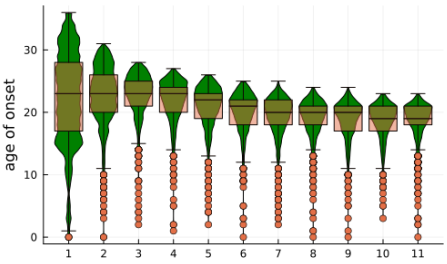
$$K(g|g^*) = Uniform(g^* - \sigma_g, g^* + \sigma_g) \tag{4.7}$$

where σ_s and σ_g are updated as the 40 quantile of the current distribution for parameters s and g [TWS+09], starting from the initial values $\sigma_s = 0.4$ and $\sigma_g = 2$.
The distance threshold is scaled considering as next value the median among all parameter vectors in the current distribution of the distances between the vector and the data, where this distance has to be defined.
The number of simulation $B_t = 1$ is kept constant trough all steps to check the results obtained in section 4.1, so the distance between a vector and the data is just the distance between LTT plots as defined in section 3.1, and the number of points in each distribution is 1000.
The number of steps is chosen in order to obtain convergence of the approximation of the posterior distribution.

The approximation of the marginal posterior distributions for all steps considered for



(a) Approximation of marginal posterior distribution for parameter s in sequential steps of ABC-SMC.

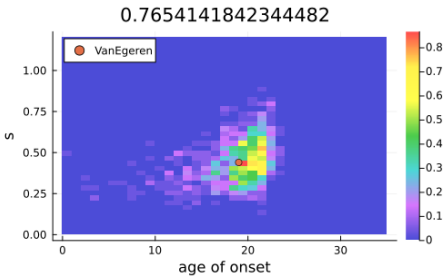


(b) Approximation of marginal posterior distribution for parameter t' in sequential steps of ABC-SMC.

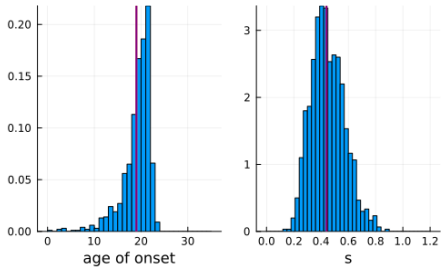
Figure 4.14: Results for ABC-SMC for patient ET2: the orange boxplots include data between the first and the third quantile, reports the median and the outliers as orange points.

both parameters are shown in figure 4.14. It is important to notice the change in the distributions over different steps depending on the sequential approximation and in particular to see that for both parameters the distributions in the last steps do not change much, meaning that convergence is reached. In figure 4.15 are presented the approximations of the posterior distributions for patient ET2 and the median values are in agreement with what obtained in [VEEN+21] and consistent with the results from section 3.1. In particular the last value obtained for the distance threshold between the simulations and the data LTT plots is $\epsilon_T = 0.765$: this suggests that a lower distance from the data can be reached.

In figure 4.16 is reported the behaviour of the distance threshold and of the cumulative

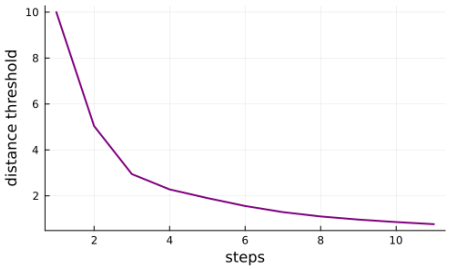


(a) Approximation of joint posterior distribution for parameters s and t' for last step of ABC-SMC. The orange point reports value from [VEEN+21] and it is reported the threshold value at last step.

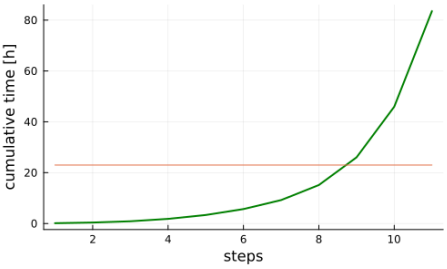


(b) Approximation of marginal posterior distribution for parameters s and t' for last step of ABC-SMC. Purple lines show the median values of the distribution, which are superimposed to the values obtained by [VEEN+21]

Figure 4.15: Approximation of posterior distribution for the last step of ABC-SMC for patient ET2.



(a) Value of the distance threshold as a function of the number of steps performed in ABC-SMC on patient ET2.

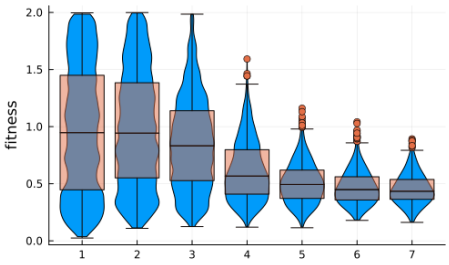


(b) Value of cumulative time in hours as a function of the number of steps performed on ABC-SMC on patient ET2, in orange is the cumulative time in hours need to run classical ABC with distance threshold 1.

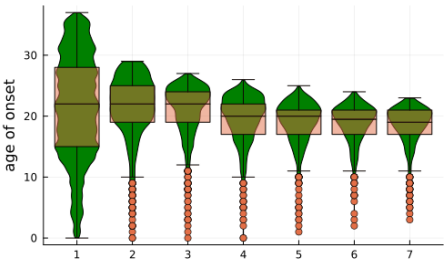
Figure 4.16

time in hours needed to perform the computations as a function of number of steps performed in ABC-SMC. In the last two steps the distance threshold decreases a little (from 0.85 to 0.76) while the time needed for computation increases dramatically (from around 40 hours to 80). This allows to state that computations could have stopped at step 10 and still a good approximation would have been obtained.

As explained in section 3.3 for each parameter vector sampled from the previous distribution and perturbed more simulations can be performed and the number of accepted ones b_t is included in the definition of the weight. To test this for patient ET2, for each parameter vector $\theta = (s, g)$, $B_t = 10$ LTT plots are produced and their distance from the experimental one compared with the threshold distance ϵ_t to be included or not in the next distribution and to calculate its weight; the distance of one parameter vector from the data is then defined as the minimum distance across B_t realizations. Since then the median of these distances across all parameter vectors in the distribution is used to update the value for the threshold in the next step, taking the minimum ensures that the sequence of distance thresholds would always be decreasing. In figure 4.17 the progression of the approximate posterior distributions for both parameters are reported, again the distributions for the last two steps do not display any significant change so it is safe to say that convergence is reached.



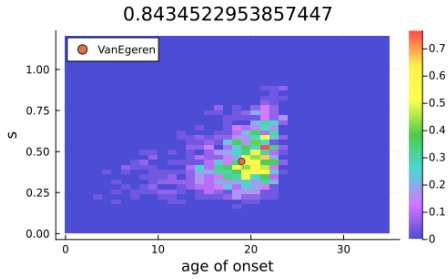
(a) Approximation of marginal posterior distribution for parameter s in sequential steps of ABC-SMC with $B_t = 10$ on patient ET2.



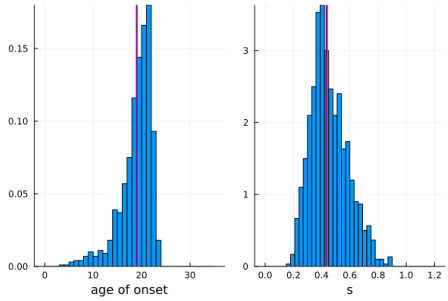
(b) Approximation of marginal posterior distribution for parameter t' in sequential steps of ABC-SMC with $B_t = 10$ on patient ET2.

Figure 4.17: ABC-SMC for patient ET2: the orange boxplots include data between the first and the third quantile, reports the median and the outliers as orange points.

The results for the approximate posterior distribution obtained for the last step are re-



(a) Approximation of joint posterior distribution for parameters s and t' for last step of ABC-SMC with $B_t = 10$ on patient ET2. The orange point reports value from [VEEN+21] and it is reported the threshold value at last step.

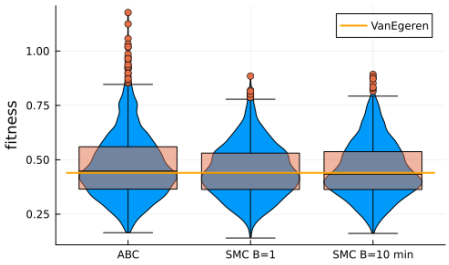


(b) Approximation of marginal posterior distribution for parameters s and t' for last step of ABC-SMC with $B_t = 10$ on patient ET2. Purple lines show the median values of the distribution and orange lines are results from [VEEN+21]

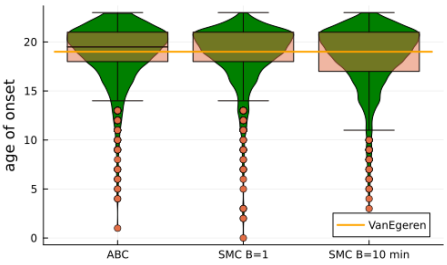
Figure 4.18: Approximation of posterior distribution for the last step of ABC-SMC with $B_t = 10$.

ported in figure 4.18 and they are consistent with what found before. It can be useful to notice that the last value for the threshold distance in this case is $\epsilon_T = 0.843$, higher than what found with $B_t = 1$.

It is possible to compare the approximation of the posterior distributions obtained with three different methodologies as shown in figure 4.19: classical ABC with $\epsilon = 1$, last step of ABC-SMC with $B_t = 1$ and last step of ABC-SMC with $B_t = 10$. The results are coherent between themselves and with what obtained in [VEEN+21] and are particularly accurate for ABC-SMC with $B_t = 1$. From figure 4.20 is is possible to notice that ABC-SMC with $B_t = 1$ gives faster results for the same value of the threshold considered compared with ABC-SMC with $B_t = 10$, which is expected since for every parameter vector only one LTT plot is produced instead of ten. Also ABC-SMC is not faster than classical ABC with similar threshold but allows to determine the right compromise between accuracy given by a lower threshold and a reasonable computation time, since the value of the threshold is not chosen a priori.



(a) Approximation of posterior distribution for parameter s for classical ABC and for last steps of ABC-SMC with $B_t = 1$ and $B_t = 10$.



(b) Approximation of posterior distribution for parameter t' for classical ABC and for last steps of ABC-SMC with $B_t = 1$ and $B_t = 10$.

Figure 4.19: Comparison of approximation of posterior distribution for different ABC methods on patient ET2, the orange boxes report values between fist and third quantiles and the median while outliers are reported as orange points. In yellow are reported values obtained by [VEEN+21]

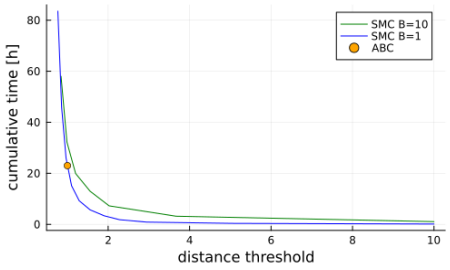


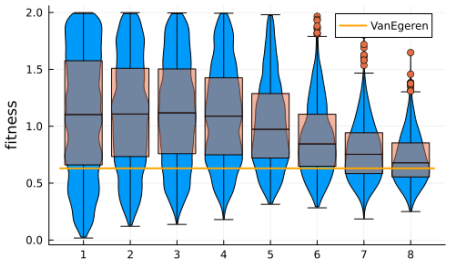
Figure 4.20: Cumulative time in hours needed to perform ABC-SMC calculations as a function of the distance threshold reached for $B_t = 10$ in green and $B_t = 1$ in blue, for classical ABC the value is reported in orange.

4.6 ABC-SMC on different patients

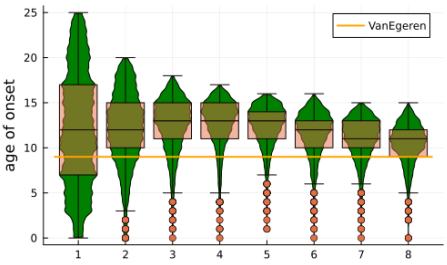
Since from figure 4.20 is possible to see that the ABC-SMC considering $B_t = 1$ allows to reach a lower value for the distance threshold while still being feasible in terms of time so it is useful to use it to infer parameters for other patients and verify the results from sections 4.1 and 4.2.

First is important to run ABC-SMC on patient ET1 to confirm the results obtained, in particular to validate the choice of threshold distance. The last value of threshold distance reached in this case is $\epsilon_T = 0.99$ which makes the results comparable with ones from 3.1. The results across steps are shown in figure 4.21 and the median values obtained in the last step are $s = 0.67(0.41 - 1.11)$ and $t' = 11(6 - 14)$. The results of the inference for SMC steps for patient PD7271 are shown in figure 4.22 and the median values of the approximation of the posterior distribution are $s = 0.98(0.74-1.25)$ and $t' = 2(1-3)$. The results for the age of onset is coherent with the confidence interval given by [WLM+22] and the last value of the threshold distance reached is $\epsilon_T = 0.413$ which is lower than the previous value used in section 4.2, it is therefore obtained a better estimation of the posterior distribution.

The results of running the ABC-SMC procedure on patient PD5163 reported in figure 4.23 are obtained reaching a minimum value of the threshold distance at $\epsilon_T = 0.211$, which is smaller that what used in section 4.2. Even tough the threshold reached is smaller and the distributions seem to have converged, since they do not seem to change much in the last steps, for parameter s in particular the approximation of the posterior distribution

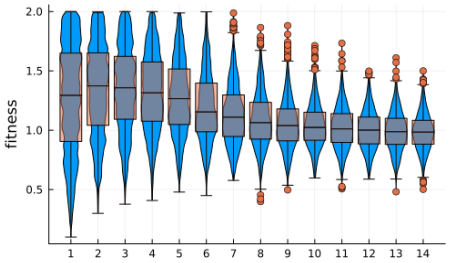


(a) Approximation of marginal posterior distribution for parameter s across steps of ABC-SMC with $B_t = 1$, orange line report value obtained from [VEEN+21].

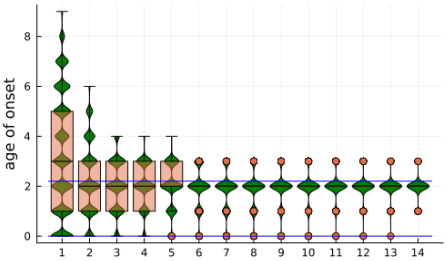


(b) Approximation of marginal posterior distribution for parameter t' across steps of ABC-SMC with $B_t = 1$,orange line report value obtained from [VEEN+21].

Figure 4.21: Results of approximation of posterior distribution across steps for ABC-SMC on patient ET1, orange boxes report values of fist and third quantile and the median, while outliers of the distributions are reported as orange points.



(a) Approximation of marginal posterior distribution for parameter s across steps of ABC-SMC with $B_t = 1$.



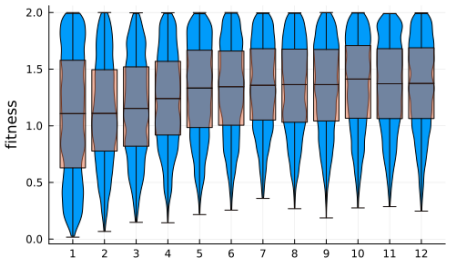
(b) Approximation of marginal posterior distribution for parameter t' across steps of ABC-SMC with $B_t = 1$, blue lines report the credibility interval for the age of onset from [WLM+22].

Figure 4.22: Results of approximation of posterior distribution across steps for ABC-SMC on patient PD7271, orange boxes report values of fist and third quantile and the median, while ouliers of the distributions are reported as orange points.

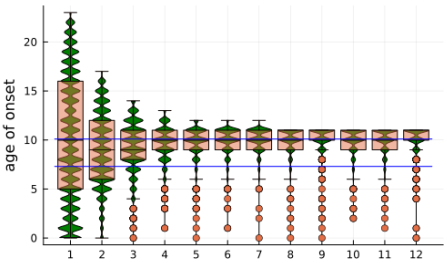
is not well defined but it seems to reproduce the uniform prior distribution. The median values obtained are similar to the ones with classical ABC $s = 1.37(0.66 - 1.93)$ and $t' = 10(8 - 11)$. In this particular case even with a lower threshold the result obtained is not the one expected, in particular for the distribution of the fitness.

4.7 Testing robustness of method on patient ET2

To test how the inference procedure can be adapted and how much it is robust with respect to variation of the data one possibility is to remove some of the data and test the model on it. In practise it is possible to remove some mutated cells from the experimental phylogenetic tree and LTT plot and perform the ABC inference on these data to see if the results are the same even with less cells sampled.
This could be useful to quantify the uncertainty of the inference, especially when considering different patients with different number of mutated sampled cells.
To to this it is considered the classical ABC performed on patient ET2 considering 13 (original number), 11, 9 and 6 mutated cells with the same parameter specification as reported in section 4.1.



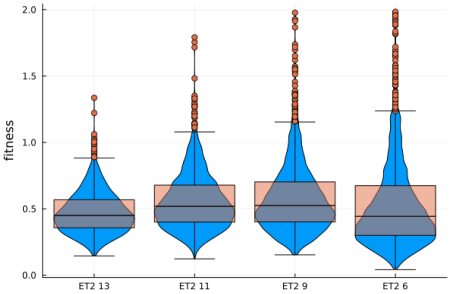
(a) Approximation of marginal posterior distribution for parameter s across steps of ABC-SMC with $B_t = 1$.



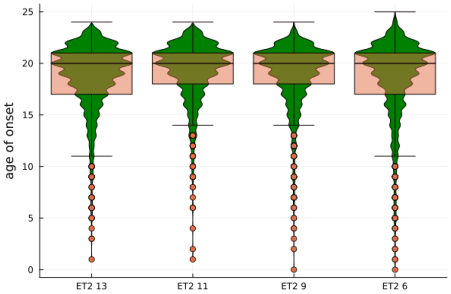
(b) Approximation of marginal posterior distribution for parameter t' across steps of ABC-SMC with $B_t = 1$, blue lines report the credibility interval for the age of onset from [WLM+22].

Figure 4.23: Results of approximation of posterior distribution across steps for ABC-SMC on patient PD5163, orange boxes report values of fist and third quantile and the median, while ouliers of the distributions are reported as orange points.

The results are reported in figure 4.24. For the age of onset the number of cells do no



(a) Approximated posterior distribution for the parameter s for different number of mutated cells considered in the data.



(b) Approximated posterior distribution for the age of onset of the mutation for different number of mutated cells considered in the data.

Figure 4.24: Approximation of the posterior distribution for parameters s and t' of patient ET2 with different number of mutated cells: the box plot in orange reports the values between the fist and the third quantile, the median and the whiskers while the outliers are reported as orange dots.

influence much the result of the inference, in particular the median does not change for each case. Instead the number of mutated cells considered influences the inference of the fitness parameter s: with decreasing number of cells the distribution is broader and so also the median value shifts. It is important to notice that, especially for the parameter fitness, the approximation of the posterior distribution is broader when less cells are considered, while still being consistent.

Chapter 5

Discussion

5.1 Summary of main contributions

In hematopoiesis the onset of the mutation JAK2-V617F in a hematopoietic stem cell is correlated with the appearance of the disease myeloproliferative neoplasm [MM17], so reconstructing the dynamics of the population of stem cells carrying that particular mutation can give insights about the development of the disease across the lifetime of an individual. It is not feasible to directly observe the mechanism regulating hematopoiesis, so it is important to develop a mathematical model to describe it.
From Single Nucleotide Variants data from patients with myeloproliferative neoplasm is reconstructed a phylogenetic tree, defining genealogical relationships between a sample of cells and how these cells developed through time, assuming a constant rate of accumulation of mutations per year.

Following what has been done in [VEEN+21] a Wright-Fisher model for cell division and renewal is studied, considering two main parameters: the fitness of the mutated cells s and the number of generations for which they develop g, which is related to the age of onset of the mutation t' as $t' = age\ of\ patient - g$. Due to its simplicity the model allows to simulate trajectories concerning the number of mutated cells through time; keeping the model simple is important when describing biological processes because it allows meaningful and clearly interpretable results without the need for extensive calculations, in additions assumptions on the process can be made explicit and then relaxed if needed. From the simulated number of mutated stem cells at each generation, phylogenetic trees and LTT plots are produced to be compared with experimental ones. This model and the process of building simulated trees introduce many sources of stochasticity, which are in principle hard to track while still allowing to capture details about the biological process.

To infer patient-specific parameters of the considered model we used an Approximate Bayesian Computation scheme as done in [VEEN+21], in which the computation of the likelihood as in classical Bayesian approach is substituted by the comparison up to a certain threshold of simulated data with experimental ones. The quality of this method, so the quality of the approximation of the posterior distribution, relies on the choice of the distance threshold ϵ which can be not well defined for all patients considered. Also since the summary statistic considered, the LTT plot, is the result of many different random processes, it is not clear how these could influence the estimation of just the two parameters s and t'.
The results of [VEEN+21] on the considered patients are reproduced, while the results we obtain with the same approach are not perfectly in agreement with [WLM+22] on their patients. Still, the results are coherent, in particular in predicting the age of onset of

the mutation years before the age of diagnosis. These particular cases could highlight the need to use a different summary statistic of the data, for example, the clonal fraction of mutated cells, to be used when comparing the simulations with experimental data, or the need to extend the Wright-Fisher model to include some features described in the different model used in [WLM+22].

In addition to what has been done before, a different definition of distance between summary statistic of the data and simulations coming from a given a parameter vector can be exploited for approximating the likelihood, in particular considering the average distance between a given number of simulations and the data. In this case, the parameters for a given patient obtained from the inference are different from the previous ones, which signals that the two distances do not approximate the likelihood of the parameters given the data in the same way. In general, this method gives a peaked approximation of the posterior distribution for the age of onset while a less defined one for the fitness, so the quality of the approximations for the two methods has to be compared to determine the best approximation.

In order to refine what obtained before is implemented a Sequential Monte Carlo version of ABC [TWS+09], which does not rely on the definition a priori of a certain threshold for the distance. The results of this method confirm previous ones and indeed refine them in the case of patients from [VEEN+21], by reaching a lower threshold distance and giving in general more defined and peaked approximate posterior distributions. In the studied case there is not a significant improvement in terms of computation time by performing ABC-SMC, but has to be taken into account that no variation in the definition of perturbation kernel is made and the number of simulations considered for the same parameter vector is maintained constant across steps; the optimization of these details could give faster results by increasing the acceptation rate of sampled parameters.

To test the quality of the results, we perform the method on the same data with fewer sampled cells considered. We found that the number of cells does not change qualitatively the approximation of the posterior distribution and in particular median values are consistent across realizations with different numbers of cells. The main difference found is that the approximation of the posterior distribution is more spread when less cells are considered, mainly regarding the fitness parameter.

5.2 Limitations of the approach and future prospective

The results found in general are coherent with what expected, highlighting how the mathematical model can allow to describe the evolution of a population of mutated stem cells in an individual and to infer the relative parameters based on patients' data. The mathematical model we used relies on assumptions made on the biological process which influence the description and final results. It is assumed that all mutated HSCs in one generation are equivalent and independent from each other, especially no death rate for cells is taken into account and it is considered that HSCs only divide symmetrically once every year. Moreover, in this case, only discrete time is considered and the number of total HSCs considered is kept constant, so in general the model can be improved to reflect the reality of the dynamics better.

The inference process is carried out in an approximate way, which relies a lot on the choice of summary statistics of the data and distance between simulation and experimental data. It is worth exploring other definitions of distance, such as the average distance between LTT plots as proposed. In particular, we are currently investigating the quality of the approximation carried out using the average distance and how much the choice of

distance influences the results and in which way. In particular, we show that for the specific case, the quality of the results of synthetic data is lower when considering as distance the average distance between multiple realizations of the LTT plot with the same parameters. The test performed on synthetic data is also a way to determine the precision of the inference depending on the parameters, which, for the considered case, seems to be lower for lower values of the fitness. It has to be noticed that the score defined does not take into account the overall quality of the distribution, i.e. the variance or if it is peaked around a single value, but only its result related to the inclusion of the true value into the credibility interval. This process has shown in addition that different quality of the results can be obtained for the two different parameters considered.
Moreover, the best results on synthetic data in terms of consistency with the true value are obtained when combining all distributions into one, which includes multiple realizations of the LTT plot for the same parameters. This approach seems to correct for the variability that arises from the stochasticity in the simulations and suggest that a population approach could give a more robust inference.
In addition, the only summary statistic of the data considered is the LTT plot, which, in general, could not be sufficient to fully describe the data since it gives no information about actual genealogical relationships between cells and the population size of mutated HSCs. Other quantities can be included, such as the clonal fraction of mutated cells. Also it is assumed that the rate of accumulation of mutations in cells in time is patient-specific and constant trough life, which disregards the possibility of the accumulation of mutations to be influenced by the patient's ambient and lifestyle (being exposed to UV rays, smoking, ecc) [YGLS+20] or in general to change in different stages of life (considering for example a higher accumulation of mutations in fetal life [WLM+22]). This main assumption influences the model and the simulations, so it could affect the results.

Given that numerous patients are considered it could be interesting to explore the possibility of introducing a population effect in the inference procedure, based on the assumption that across individuals the parameters to be estimated can be seen as sampled from a common unknown distribution. The hyperparameters of this unknown distribution can also be estimated in a Bayesian setting, like exploiting a Hierarchical Bayesian approach [TZ14], and they would be inferred based on many patients' data. This could decrease the risk of overfitting the parameters to a given patient data and correct the inference on a specific patient to be more coherent with the population distribution. In addition when considering practical applications, for example in diagnostics, describing the dynamics of mutated cells has to be done at the population level.

Bibliography

[BCMR09] Mark A. Beaumont, Jean-Marie Cornuet, Jean-Michel Marin, and Christian P. Robert. Adaptive approximate bayesian computation. *Biometrika*, 96, 2009.

[CMR+21] Tim H. H. Coorens, Luiza Moore, Philip S. Robinson, Rashesh Sanghvi, Joseph Christopher, James Hewinson, and Moritz J. Przybilla et al. Extensive phylogenies of human development inferred from somatic mutations. *Nature*, 597, 2021.

[ELP+05] Baxter EJ, Scott LM, Campbell PJ, East C, Fourouclas N, Swanton S, Vassiliou GS, Bench AJ, Boyd EM, Curtin N, Scott MA, Erber WN, and Green AR. Acquired mutation of the tyrosine kinase jak2 in human myeloproliferative disorders. *Lancet*, 365, 2005.

[Fis23] R. A. Fisher. Xxi.âon the dominance ratio. *Proceedings of the Royal Society of Edinburgh*, 42, 1923.

[GF20] Clara Grazian and Yanan Fan. A review of approximate bayesian computation methods via density estimation: Inference for simulator-models. *WIREs Computational Statistics*, 12(4):e1486, 2020.

[HRB+22] Gurvan Hermange, Alicia Rakotonirainy, Mahmoud Bentriou, Amandine Tisserand, Mira El-Khoury, FranÃ§ois Girodon, Christophe Marzac, William Vainchenker, Isabelle Plo, and Paul-Henry CournÃ¨de. Inferring the initiation and development of myeloproliferative neoplasms. *Proceedings of the National Academy of Sciences*, 119, 2022.

[JCEea13] Nangalia J, Massie CE, Baxter EJ, and et al. Somatic calr mutations in myeloproliferative neoplasms with nonmutated jak2. *N Engl J Med*, 369, 2013.

[JGD+06] Catriona H. Jamieson, Jason Gotlib, Jeffrey A. Durocher, Mark P. Chao, M. R. Mariappan, Marla Lay, Carol Jones, James L. Zehnder, Stan L. Lilleberg, and Irving L. Weissman. The jak2 v617f mutation occurs in hematopoietic stem cells in polycythemia vera and predisposes toward erythroid differentiation. *Proceedings of the National Academy of Sciences*, 103, 2006.

[LSOS+18] Henry Lee-Six, Nina Friesgaard Obro, Mairi S. Shepherd, Sebastian Grossmann, Kevin Dawson, Miriam Belmonte, and Robert J. Osborne et al. Population dynamics of normal human blood inferred from somatic mutations. *Nature*, 561, 2018.

[MM17] Adam J. Mead and Ann Mullally. Myeloproliferative neoplasm stem cells. *Blood*, 129, 2017.

[MMPT03] Paul Marjoram, John Molitor, Vincent Plagnol, and Simon TavarÃ©. Markov chain monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100, 2003.

[SN87] N Saitou and M Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4, 1987.

[TWS+09] Tina Toni, David Welch, Natalja Strelkowa, Andreas Ipsen, and Michael P.H. Stumpf. Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 6, 2009.

[TZ14] Brandon M. Turner and Trisha Van Zandt. Hierarchical approximate bayesian computation. *Psychometrika*, 79, 2014.

[VEEN+21] Debra Van Egeren, Javier Escabi, Maximilian Nguyen, Shichen Liu, Christopher R. Reilly, Sachin Patel, and Baransel Kamazet al. Reconstructing the lineage histories and differentiation trajectories of individual cancer cells in myeloproliferative neoplasms. *Cell Stem Cell*, 2021.

[VK17] William Vainchenker and Robert Kralovics. Genetic basis and molecular pathophysiology of classical myeloproliferative neoplasms. *Blood*, 129, 2017.

[WLL+12] John S. Welch, Timothy J. Ley, Daniel C. Link, Christopher A. Miller, David E. Larson, Daniel C. Koboldt, and Lukas D. Wartman et al. The origin and evolution of mutations in acute myeloid leukemia. *Cell*, 150, 2012.

[WLM+22] Nicholas Williams, Joe Lee, Emily Mitchell, Luiza Moore, E. Joanna Baxter, James Hewinson, and Kevin J. Dawson et al. Life histories of myeloproliferative neoplasms inferred from phylogenies. *Nature*, 602, 2022.

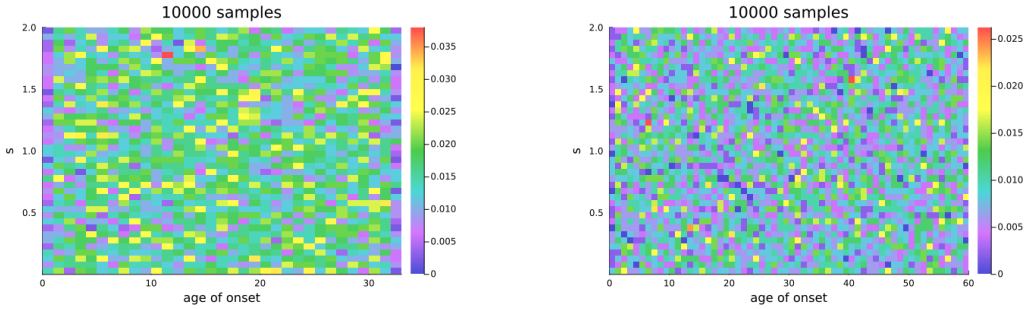
[Wri31] Sewall Wright. Evolution in mendelian populations. *Genetics*, 16, 1931.

[YGLS+20] Kenichi Yoshida, Kate H. C. Gowers, Henry Lee-Six, Deepak P. Chandrasekharan, Tim Coorens, Elizabeth F. Maughan, and Kathryn Beal et al. Tobacco smoking and somatic mutations in human bronchial epithelium. *Nature*, 578, 2020.

Appendix A

Approximation of prior distribution

To test the accuracy of the algorithm and to ensure that the sampling is correctly done from a uniform prior distribution, at first the inference algorithm is tested without considering the cut-off distance epsilon and so retaining every value in the distribution. It is expected to obtain again a uniform distribution of the parameters because the algorithm would not reject any parameters vector. In figure A.1 it is possible to see the approximations for



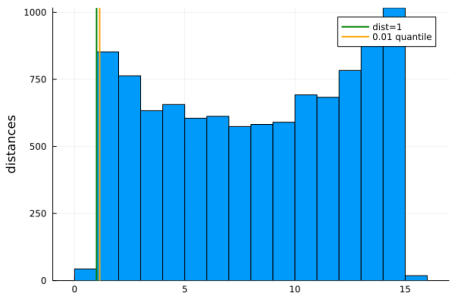
(a) Approximation of uniform prior distribution for parameters s and t' in range considered for patient ET1

(b) Approximation of uniform prior distribution for parameters s and t' in range considered for patient ET2

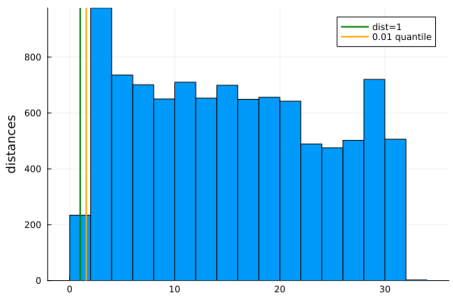
Figure A.1

posterior distributions for the two patients without considering distance, which reproduce the uniform distribution of the parameters as expected and shows that at least thousands of points has to be sampled in order to sample from the whole prior distribution.

In figure A.2 are reported the histograms showing the distribution of distances between the simulated data and the experimental ones relative to both patients. Since the cutoff distance used $\epsilon = 1$ is lower than the 0.01 quantile in both cases, it means that in the approximation of the posterior distribution less than 1% of the sampled parameter vectors would be retained.



(a) Histogram with distribution of distances between simulation and experimental data of patient ET1, the green line correspond to the value $\epsilon = 1$ used at first in the ABC procedure and the orange line is the 0.01 quantile



(b) Histogram with distribution of distances between simulation and experimental data of patient ET2, the green line correspond to the value $\epsilon = 1$ used at first in the ABC procedure and the orange line is the 0.01 quantile

Figure A.2