

Raúl Sanz Ruiz

Inferring Tumor Growth Models from Variant Allele Frequencies

MASTER'S THESIS

Supervisor: Dr. Donat Wenghorn (Centre for Genomic Regulation)

Tutor: Dr. Mireia Olivella (Universitat Rovira i Virgili)

Master's Degree in Health Data Science



Barcelona, 2025

Index

1. Introduction	1
1.1. Background and motivation	1
1.1.1. Cancer through the lens of population genetics	1
1.1.2. Next-generation sequencing	3
1.1.3. Variant allele frequencies	4
1.2. Objectives	6
1.3. The state-of-the-art	7
2. Methods	9
2.1. Dataset	9
2.2. Preprocessing	10
2.3. Tumor Allele Frequency Interpreter	11
2.3.1. VAF reconstruction	12
2.3.2. Approximate Bayesian Computation	12
2.3.3. The loss function	13
2.3.4. Model classification	14
2.4. Validation	15
2.5. Downstream analyses	16
2.5.1. Differential gene expression analysis	16
2.5.2. Survival analysis	17
3. Results	18
3.1. Data exploration	18
3.2. Predictions	19
3.2.1. On simulations	19
3.2.2. On real data	20
3.3. Downstream analyses	23
3.3.1. Differential gene expression	23
3.3.2. Survival analysis	26
3.4. TAFI vs. MOBSTER	28
4. Discussion	30
5. Conclusions	32
5.1. Limitations	32
5.2. Future work	32
5.3. Ethical and diversity aspects	33
Code Availability	35
References	35

1. Introduction

1.1. Background and motivation

1.1.1. Cancer through the lens of population genetics

Cancer development can be seen as an evolutionary process. From its initial stage as a single aberrant cell to its progression into a heterogeneous tumor, cancer evolves through the accumulation of genetic mutations and the selective expansion of the ones that acquire a fitness advantage (Greaves & Maley, 2012) [1]. This evolutionary dynamic can be studied using the principles of population genetics, a field originally developed to study how genetic variation in organisms change over time under the influence of evolutionary forces. While classical population genetics focuses on changes across generations of organisms, its mathematical principles can be directly applied to cancer by defining a tumor as a population of cells undergoing somatic evolution (Beerenwinkel et al., 2015) [2].

In traditional population genetics, mutation serves as the ultimate source of genetic variation, introducing new alleles into a population. In cancer, mutations occur somatically due to replication errors, environmental factors, or defective DNA repair mechanisms (Martincorena & Campbell, 2015) [3]. These mutations can be mainly categorized into driver mutations—which confer a selective advantage by enhancing proliferation, survival, or metastatic potential—or neutral mutations—which accumulate without significantly impacting fitness (Vogelstein et al., 2013) [4]. The interplay between these types of mutations shapes the tumor’s genetic architecture over time.

Starting from a single cell that has acquired driver mutations for tumorigenesis—the process by which normal cells transform into cancerous ones—the resulting population of cells begins to expand. Selection then acts upon the genetic variation arising within this growing population, increasing the frequency of beneficial alleles while eliminating deleterious ones. This process is also defined as clonal expansion, where subpopulations carrying driver mutations outcompete others, leading to selective sweeps: rapid expansions of genetically distinct clones (McFarland et al., 2013) [5]. On the other hand, genetic drift—the random fluctuation of allele frequencies—plays a critical role in early tumor growth or metastatic seeding, where small populations of cells may experience stochastic loss or fixation of mutations independent of selective effects (Bozic et al., 2010) [6].

One of the foundational models in population genetics is the Wright-Fisher model (Fisher, 1930; Wright, 1931) [7][8], which provides a theoretical framework for understanding how allele frequencies fluctuate under the influence of genetic drift. This model assumes a fixed population size, discrete non-overlapping generations, and

random sampling of individuals to form the next generation. In the absence of mutation and selection, the Wright-Fisher model describes purely stochastic changes in allele frequencies, referred to as neutral evolution. When applied to cancer, this model offers a baseline for understanding how passenger mutations, which do not confer a fitness advantage or disadvantage, accumulate over time due to random sampling during cell divisions. Such a neutral process is especially relevant in tumors that evolve slowly or lack strong selective pressures (Williams et al., 2016) [9].

However, many tumors do not follow neutral dynamics. In rapidly growing tumors, non-neutral models that incorporate exponential growth and positive selection more accurately capture the observed evolutionary behavior (Sottoriva et al., 2015) [10]. In these models, driver mutations expand rapidly and dominate the tumor cell population in a process known as a selective sweep. This alters the allele frequency distribution in a characteristic way: fewer variants are observed at intermediate frequencies, and rare variants are overrepresented.

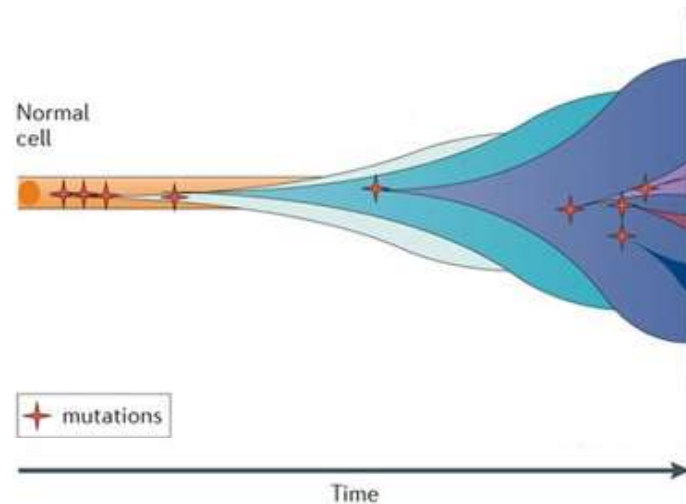


Figure 1. Evolution and expansion in tumor progression. Sequential acquisition of driver mutations (color-coded) within a population of cancer cells over time. Beginning with a non-mutated normal cell (far left), successive mutations confer selective advantages, leading to clonal expansions. Over time, certain mutations (e.g., orange, cyan, blue, and violet) become fixed in the population due to their fitness benefits, while others (e.g., purple, red, and dark-blue) that emerged more recently remain at subclonal frequencies. Adapted from: Yates & Campbell (2012) [11].

Modelling cancer with population genetics not only describes mutational patterns, but defines a quantitative framework to infer the evolutionary history of tumors. These models reconstruct clonal composition, measure selective pressures on different subclones, and determine the evolution model. For this reason, by integrating population

genetics principles with tumor sequencing data, allows us to interpret the genetic diversity seen among patients and cancer types.

1.1.2. Next-generation sequencing

Since the completion of the Human Genome Project (International Human Genome Sequencing Consortium, 2004) [12], Next-Generation Sequencing (NGS) technologies have transformed genomics by offering faster, more accurate, and cost-effective alternatives to traditional methods like Sanger sequencing (Metzker, 2010) [13]. These advancements have revolutionized our ability to study genetic variation, understand disease mechanisms, and develop personalized medicine.

NGS allows the parallel sequencing of millions of DNA or RNA molecules simultaneously. The process begins with fragmenting the genetic material and ligating specific adapters to the fragment ends, which are essential for sequencing (Shendure & Ji, 2008) [14]. These fragments are then amplified and immobilized on a flow cell, where they get sequenced using platforms such as Illumina, Ion Torrent, or Oxford Nanopore (van Dijk, et al., 2014)[15]. While NGS has transformed high-throughput sequencing, it is not perfect. Biases during fragmentation and adapter ligation can affect accuracy, so rigorous quality control is essential.

DNA sequencing is a fundamental NGS technology in genomics that enables the determination of the precise order of nucleotides within a DNA molecule. It can be broadly categorized into two main approaches: whole-genome sequencing (WGS) and whole-exome sequencing (WES). WGS aims to sequence the entire genome, including coding and non-coding regions. In contrast, WES focuses only on the exome—the protein-coding regions of the genome—which only corresponds to 1-2% of the genome but includes a high proportion of relevant (e.g. disease-associated) mutations (Choi, et al., 2009) [16]. Because of this targeted approach, WES achieves higher coverage per nucleotide in coding regions, allowing the detection of low-frequency mutations that may be missed in WGS. However, WGS's broader coverage makes it more suitable for studying structural variants, non-coding mutations, and complex traits.

A common next step that follows the sequencing is the identification of genetic variants through a process known as variant calling. This involves comparing the sequenced DNA fragments to a reference genome, which is a standardized representation of the genome for a species, serving as a baseline for comparison. The most widely used human reference genome nowadays is the GRCh38 (hg38) version, which is continuously updated to improve accuracy and completeness (Schneider, et al., 2017) [17]. Variant calling detects differences between the sequenced data and the reference genome, such as single nucleotide variants (SNVs), insertions, deletions, and structural variants.

In cancer genomics, the analysis typically focuses on identifying somatic mutations by comparing a tumor sample to a matched normal sample (often extracted from blood),

rather than a reference genome. The accuracy of variant calling relies on the quality and depth of the sequencing data, as well as the computational tools used to align sequences and analyze discrepancies. There are many tools available for variant calling, which, despite their differences, rely on similar principles of sequence alignment and statistical inference to identify variants. The output of variant calling is typically stored in a Variant Call Format (VCF) file, which includes a list of genomic positions, the identified variants, and associated metadata such as quality scores, read depths, and allele frequencies. This information is crucial for interpreting the biological significance of variants, whether in the context of disease research, population genetics, or personalized medicine.

Another NGS technique is RNA sequencing (RNA-seq), which is used to analyze gene expression by sequencing the transcriptome—the complete set of RNA molecules in a cell or tissue. Unlike DNA sequencing, which provides information about the genetic code, RNA-seq captures the dynamic expression of genes, revealing which genes are active, at what levels, and how they are regulated under specific conditions (Wang, et al., 2009) [18]. The RNA-seq workflow begins with the extraction of RNA, which is then converted into complementary DNA (cDNA) through reverse transcription. The cDNA is fragmented, adapters are added, and the fragments are sequenced using NGS platforms. The resulting data is analyzed using bioinformatics tools to align reads to a reference genome or transcriptome, quantify gene expression levels, and identify differentially expressed genes between samples or conditions.

DNA sequencing and RNA-seq are complementary methods that together provide a complete picture of genetic information and its functional expression. DNA sequencing reveals the static genetic code, identifying mutations and structural variants, while RNA sequencing captures the dynamic activity of genes, showing how they are expressed and regulated in specific conditions. This combination is critical to understand the full impact of genetic changes, such as how a mutation identified by DNA sequencing alters gene expression or cellular pathways, as revealed by RNA-seq. By integrating these approaches, it is possible to uncover deeper insights into disease mechanisms, identify therapeutic targets, and advance personalized medicine, making their combined use indispensable in modern genomics.

1.1.3. Variant allele frequencies

Variant allele frequency (VAF) is a fundamental metric in genomics that quantifies the proportion of sequencing reads supporting a specific variant relative to the total reads at a given genomic position. It is calculated using the following formula:

$$VAF = \frac{AD_{ALT}}{coverage} = \frac{AD_{ALT}}{AD_{REF} + AD_{ALT}} \quad (1)$$

where AD_{REF} (allelic depth reference) represents the number of reads supporting the reference allele, AD_{ALT} (allelic depth alternative) represents the number of reads supporting the variant allele, and coverage is the total number of reads at that position ($AD_{REF} + AD_{ALT}$). Variant allele frequencies are critical for understanding the prevalence of a variant within a sample.

While a single-variant frequency does not explain the tumor evolution itself, it is possible to extract far more information when aggregating all the variants (the result of applying variant calling on a sequenced tumor) and building a VAF distribution. A VAF distribution provides a global view of the genetic landscape of a tumor, revealing patterns that are not apparent when examining individual mutations in isolation.

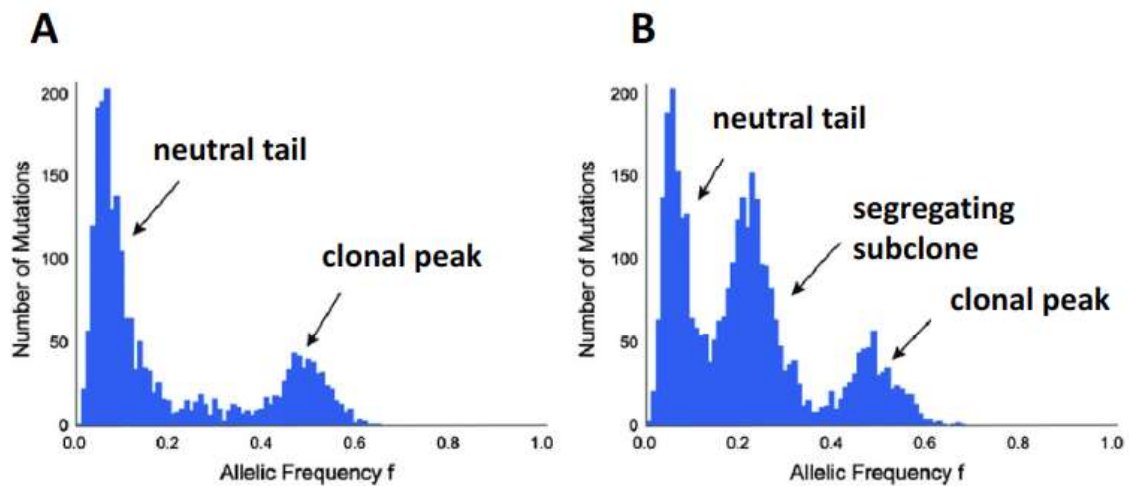


Figure 2. Shape of VAF distributions. (A) Simulated distribution of VAFs in a cancer evolving neutrally. The distribution shows a peak around 0.5, representing clonal variants present in all cancer cells. Neutral variants, present at lower frequencies, follow a $1/f$ distribution. (B) VAF distribution for a cancer in which a fitter subclone has expanded clonally. The subclone's mutation frequencies are increased during clonal expansion, creating a distinct cluster of mutations. Despite the presence of selection, a background of neutral evolution persists, reflecting the stochastic accumulation of mutations within the selected subclone and the residual tumor cells. Adapted from: Graham & Sottoriva (2017) [19].

In cancer genomics, the distribution of VAFs can be characterized by three key components, as illustrated in Figure 2:

- The clonal peak: This peak consists of mutations present in all tumor cells, appearing at a frequency determined by tumor purity and ploidy. These mutations

typically arise early in tumorigenesis —the process by which a tumor develops— and are shared across all tumor cells.

- The subclonal peak (or neutral tail): This component includes mutations that are present in only a subset of tumor cells, reflecting intratumoral heterogeneity. These mutations occur later in tumor evolution, after the initial clonal expansion, and are often the result of branching events in the tumor’s evolutionary history.
- What we denominate “Segregating subclone”: Refers to a group of mutations that co-occur in a subset of tumor cells and appear at intermediate allele frequencies indicating that the subclone has not yet been fixed but is expanding or persisting alongside other subclones or the main clone (Nik-Zainal et al., 2012) [21]. There might be different segregating subclones, and each of them corresponds to a distinct subpopulation of tumor cells that have acquired additional mutations, which may confer selective advantages or simply reflect neutral drift (McGranahan & Swanton, 2017) [20].

The VAF distribution in a tumor is importantly influenced by two key factors: tumor purity and ploidy. Tumor purity, which ranges from 0 (no tumor cells) to 1 (pure tumor), reflects the proportion of tumor cells in a sample, with lower values indicating contamination by normal or immune cells. Ploidy, on the other hand, refers to the number of chromosome sets in a cell —while diploid tumors have two sets, deviations result in aneuploidy (e.g. whole-genome duplications). The position of the clonal peak in the VAF distribution is scaled by the factor $\frac{\text{purity}}{\text{ploidy}}$. For instance, in a diploid tumor (ploidy = 2) with 100% purity (purity = 1.0), a clonal mutation is expected at a VAF of 0.5. However, if purity drops to the 60%, the expected VAF shifts to 0.3, moving the clonal peak toward lower allele frequencies. This shift can cause overlap between clonal and subclonal peaks, complicating the identification of subclonal populations.

1.2. Objectives

After discussing the importance of studying VAF distributions, we developed a statistical algorithm to infer the evolutionary history and dynamics of tumors from them: Tumor Allele Frequency Interpreter (referred to as “TAFI” from now on).

Our primary objective is to accurately reconstruct the true number of clonal (C) and subclonal (S) mutations within a tumor, as well as to estimate the tumor purity (i.e., the fraction of cancer cells within the sample). These estimates reflect the actual underlying values of these parameters, rather than simply describing the values observed from the sequencing data.

A second major goal of TAFI was to classify tumors according to their evolutionary mode. Specifically, we categorize samples into two groups, corresponding to distinct evolutionary models known from population genomics: The Wright-Fisher (“WF”) model, representing evolutionary dynamics characterized by constant population size and neutral evolution; and the exponential (“EXP”) evolution model, reflecting rapidly expanding tumors.

Following parameter inference and classification, we aim to carry out downstream analyses to explore the biological distinctions between tumors evolving under different regimes. Our working hypothesis is that EXP-classified tumors will exhibit signs of accelerated evolutionary dynamics potentially correlating with features of increased malignancy, treatment response, or poor prognosis. These analyses will serve both to validate TAFI’s classification framework and to generate insights into the adaptive strategies exploited by different tumors during progression.

1.3. The state-of-the-art

Several computational strategies have been proposed to extract evolutionary information from VAF data. These include deconvolution-based methods that cluster mutations into clonal and subclonal categories, phylogenetic algorithms that aim to reconstruct tumor trees from bulk or single-cell data, and probabilistic models that interpret the VAF histogram as the result of underlying population-level dynamics.

Within this last category highlights MOBSTER (model-based clustering in cancer) (Caravagna, et al., 2020) [22], which introduces a statistical framework to describe the evolutionary structure of a tumor based on its VAF distribution. MOBSTER uses a model that assumes that variant allele frequencies arise from a combination of distinct generative processes: uses a power-law component (Pareto distribution) to reflect the subclonal mutations expected under neutral evolution, and one or more binomial distributions to describe clonal and selectively expanding subclonal populations (segregating subclones). The underlying assumption is that the tail of the VAF distribution, containing low-frequency mutations, follows a Pareto distribution when evolution is neutral, whereas deviations from this pattern indicate selective sweeps or clonal expansions.

The key parameter estimated by MOBSTER is the “shape” parameter α of the Pareto distribution that indicates the evolutionary mode. A shape parameter close to one indicates neutral evolution, where the accumulation of subclonal mutations is largely governed by genetic drift. Higher values of α suggest deviations from neutrality, potentially due to shorter time since clone expansion. MOBSTER performs model fitting via maximum likelihood estimation, and assigns mutations to their most probable mixture component. This results in a classification of mutations into clonal,

subclonal-neutral, or subclonal-selected categories, allowing users to quantify the evolutionary heterogeneity of tumors.

Despite its conceptual strengths and interpretability, MOBSTER has several limitations. It does not infer global parameters such as the total number of clonal or subclonal mutations, or estimate tumor purity directly from the data. While it can suggest whether a tumor follows constant or exponential dynamics based on the fitted α parameter, it does not account for the sampling noise introduced by read depth and tumor purity. In addition, the VAF distribution can be significantly affected by filtering steps, which are not considered by MOBSTER but can bias the inference of the subclonal components.

2. Methods

2.1. Dataset

We assembled a dataset of 20,914 tumor samples by integrating data from three major cancer sequencing initiatives: the Pan-Cancer Analysis of Whole Genomes (PCAWG) (The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, 2020) [23], the Multi-Center Mutation Calling in Multiple Cancers (MC3) (Ellrott, et al., 2018) [24], and the Hartwig Medical Foundation (HMF) dataset (Hartwig Medical Foundation, n.d.) [25]. A summary of the key characteristics of each cohort is presented in Table 1 below.

Project	Sequencing	# Samples	Expression data
MC3	WES	10,537	Yes
PCAWG	WGS	2,697	Yes*
HMF	WGS	7,198	Yes

Table 1. Summary of dataset characteristics. The table provides an overview of each project, including the following details: 1) *project* abbreviation 2) *sequencing* technique employed 3) total number of *samples*, and 4) availability of RNA-seq *expression data*.

All three projects are pan-cancer, meaning they include tumor samples from different cancer types. To ensure consistent labeling of tumor types across datasets, we annotated the samples using TCGA study abbreviations (Supplementary Table 1). Another important variable to account for is tumor stage, which varies from sample to sample and can include stages such as primary, locally advanced, or metastatic. Notably, the HMF dataset is predominantly composed of metastatic tumors.

The raw genomic data were downloaded in Variant Call Format (VCF), the standard output of variant calling pipelines. However, since these datasets originate from different projects, they were generated using distinct sequencing platforms, bioinformatics pipelines, and variant calling strategies. These differences may affect the variant detection, potentially introducing biases when comparing cohorts. Therefore, it is crucial to understand the methodologies used in each project:

- PCAWG: Researchers independently ran three somatic SNV-calling pipelines on all samples: (1) the Sanger pipeline, which uses algorithms such as CaVEMan; (2) the DKFZ pipeline, which includes a custom SNV caller optimized for cancer genomes; and (3) the Broad pipeline, which employs MuTect2. To

enhance the reliability of variant calls, a "two-plus" consensus strategy was applied, retaining only those SNVs identified by at least two of the three pipelines in the final call set.

- MC3: Variant calling was performed using several tools, including MuTect, MuSE, VarScan2, Radia, Pindel, SomaticSniper, and Indelocator. Based on performance benchmarks reported in the corresponding publication, we built a consensus mutation set using MuTect and MuSE, which demonstrated high accuracy.
- HMF: This project used the Somatic Alterations in Genome (SAGE) pipeline, a proprietary tool optimized for high-depth, clinical-grade tumor sequencing.

** The RNA-seq data provided by PCAWG was already normalized sample-wise in TPM (transcripts per million) rather than provided as raw counts. This is not ideal for analyses that compare different samples, such as differential gene expression analysis.*

2.2. Preprocessing

As mentioned, the starting point for our analysis was the collection of VCF files provided by the projects. VCF files store a lot of information about genetic variants, but due to their complexity, accessing and locating relevant information can be challenging. To simplify this process, we first converted the VCF files into BED (Browser Extensible Data) format. BED is a simpler, tab-delimited text format used to represent genomic regions using coordinates and associated annotations.

The newly generated BED files contain only a curated subset of the data from the original VCFs —specifically, the information required for our analysis. Each entry in the BED file includes:

- Genomic coordinates: chromosome (chr), start, and end positions.
- Nucleotide information: reference allele (REF) and alternative allele (ALT).
- Allele depths: number of reads supporting the reference (AD_REF) and alternative (AD_ALT) alleles.
- Sequencing coverage (DP), representing the total number of reads at the position.
- Filter status (FILTER), indicating the quality/confidence of the variant call and whether it passed predefined quality thresholds.

Then, those files followed the pipeline and passed through what we defined as “sample selector”. This sample selector is a script that classifies the samples based on their VAF, keeping the samples that are possible to apply TAFI on them and discarding the ones where it is not. The sample selector follow the next criterion:

1. Only mutations flagged as "PASS" in the FILTER column of the variant calling are kept. This helps ensure that only high-confidence, likely true-positive mutations are considered.

2. Mutations located in genomic regions with a copy number equal to 2 are kept, aligning with TAFI's diploid genome model assumption. Variants in regions with aberrant copy number states (due to many possible reasons) are excluded, as they may distort the variant allele frequency and result in being misclassified.
3. Exclude mitochondrial mutations: Only variants located on autosomal and sexual chromosomes (chromosomes 1-22, X and Y) are kept. Mitochondrial mutations are excluded because they follow different inheritance and replication mechanisms.
4. Variants with less than three alternative reads are discarded to remove extremely low-frequency variants that may represent sequencing errors or artifacts.
5. Samples with fewer than 1,000 mutations are excluded. A low mutation count may indicate insufficient data for accurate modeling and could lead to unreliable TAFI inference.
6. For each remaining sample, the variant allele frequency is computed following the formula 1 previously described. The distribution of VAFs is then evaluated using a Kernel Density Estimation (KDE) approach, implemented via the `gaussian_kde` function from the Python `scipy.stats` package (version 1.15.2). Then:
 - i. Samples with more than two peaks in the KDE-estimated VAF distribution are excluded, as TAFI is currently designed to handle only simple clonality patterns (with one or two peaks, corresponding to clonal and subclonal) and not samples containing segregating subclones.
 - ii. Samples where the clonal peak is centered at a frequency larger than 0.5 are also discarded. In diploid organisms, a VAF above 0.5 implies a purity greater than 1, which is not possible and may indicate underlying issues such as copy number aberrations or misannotation.

All samples that passed the sample selector were considered valid candidates for TAFI.

2.3. Tumor Allele Frequency Interpreter

We developed TAFI (abbreviation of “Tumor Allele Frequency Interpreter”), an iterative statistical algorithm designed to infer key parameters of tumor evolution by modeling and fitting VAF distributions.

The core idea behind TAFI is to infer biologically meaningful parameters by simulating the sequencing process under different evolutionary scenarios and comparing the resulting VAF distributions to those observed in real tumor data. By systematically exploring combinations of parameters, TAFI identifies the configurations that best reproduce the empirical VAF distribution, enabling robust inference of tumor purity, clonal structure, and evolutionary dynamics.

2.3.1. VAF reconstruction

At each iteration, TAFI reconstructs a simulated VAF distribution based on a candidate set of parameters, aiming to replicate the sequencing process to be comparable to the empirical VAF distribution observed in a tumor sample. The parameter set includes the tumor purity, the number of clonal (C) and subclonal (S) mutations, and the evolutionary model assumed for the tumor, which can be either Wright-Fisher (WF) or Exponential (EXP). Additionally, the simulation needs the empirical distribution of coverage values from the sequencing data, which is donor-specific.

To generate a simulated VAF distribution, TAFI proceeds in three main steps.

First, the clonal mutations are simulated. This step is identical across both evolutionary models. For each of the C clonal mutations, coverage values are sampled from the original coverage distribution observed in the tumor sample. Then, for each coverage value, the number of reads supporting the alternative allele is drawn from a binomial distribution, where the success probability is modeled by the tumor purity and ploidy.

The second step involves simulating the subclonal mutations, which are also sampled using a binomial distribution. However, unlike the clonal ones, the success probabilities for subclonal mutations vary depending on the assumed evolutionary model. In the Wright-Fisher model, the frequencies of these mutations follow a Pareto distribution (equation 2) with shape parameter $\alpha = 1$, consistent with neutral evolution. In the exponential model, the same distribution is used but with $\alpha = 2$, reflecting a steeper decay. This difference in α directly modulates the likelihood of observing subclonal mutations at different frequencies, resulting in a more pronounced skew toward low-frequency mutations under the exponential growth model. The success probability of those mutations is also modeled by purity and ploidy. (see Formula 2)

$$f(\alpha) = \frac{1}{x^\alpha} \times \frac{\text{purity}}{\text{ploidy}} \quad (2)$$

The final step involves computing the VAF values from the simulated read counts by dividing the number of alternative reads over the total coverage at that site.

2.3.2. Approximate Bayesian Computation

Once we established a method for simulating VAF distributions from a given set of parameters, the next objective was to identify the optimal configuration that produces the simulation most closely matching the original observation —denoted as $\theta = \{\text{purity}, S, C\}$. To achieve this, we employ Approximate Bayesian Computation (ABC), a likelihood-free inference method that works by iteratively simulating data under different parameter combinations and comparing the results to the empirical data

using a distance or loss function. Parameters that generate simulated data that closely replicate the observed data are retained, while others are discarded or adjusted.

In our implementation, we initialize the inference process with 100 randomly sampled starting points, each representing a unique combination of parameters (θ') drawn from biologically plausible ranges. For each of these starting points, a simulated VAF distribution is generated using the reconstruction procedure described earlier. The similarity between the simulated and observed VAF distributions is evaluated using a custom loss function, defined in detail in the following section.

Then, the algorithm proceeds to explore the parameter space by proposing new candidate values for the three parameters from each starting point. This is done by introducing small random perturbations to the current values: tumor purity is modified by adding Gaussian noise, while the values of S and C are adjusted by adding normally distributed random fluctuations scaled to their magnitudes. These new candidate values are then used to simulate a new VAF distribution, which is again compared to the observed data. If the new set of parameters produces a smaller loss (i.e., better fit), it replaces the previous one; otherwise, the original parameters are retained. This stochastic search process is repeated over 100 iterations for each of the 100 starting points, allowing to refine its estimate of the parameters that best explain the observed VAF distribution.

Our default design explores a 100×100 search space, resulting in a total of 10,000 simulations being evaluated per run. However, this resolution is configurable depending on the use case. Increasing the number of starting points or iterations may improve the precision of the inference at the cost of increasing the computational time. Conversely, for applications requiring faster inference or lower computational resources, the search space can be reduced to achieve a trade-off between speed and accuracy.

2.3.3. The loss function

As in any optimization problem, a loss function is required to evaluate each iteration and determine the optimal set of parameters that minimize it. The loss function used in our method is defined as:

$$Loss(\theta) = \frac{wasserstein(VAF_{obs}, VAF_{sim(\theta)})}{wasserstein_{max}(VAF_{obs})} + \lambda * \frac{|N_{obs} - N_{sim(\theta)}|}{N_{obs}} \quad (3)$$

This equation aims to quantify the discrepancy between the observed VAF (derived from real data) and the simulated VAF (generated under a given set of parameters).

The first term of the equation, the Wasserstein distance (also called the Earth Mover's Distance), quantifies the difference between two probability distributions—the observed and simulated VAFs in this case— by measuring the minimum "work" required to transform one into the other. Mathematically, for two discrete distributions, the Wasserstein distance is computed as the sum of the absolute differences between their sorted cumulative distributions. Unlike other metrics (e.g., sum-of-squares), it does not require binning, making it more robust for comparing continuous distributions.

However, a key limitation is that it does not account for the absolute number of mutations (N), focusing solely on distribution shape. To address this limitation, we introduced the absolute relative difference between the observed and the simulated mutation counts as the second term in the function, ensuring that the optimization process considers both the shape and scale of the VAF.

The challenge was then weighting the two terms appropriately to prevent one from dominating the other. If the Wasserstein term is weighted too heavily, the optimization may yield a VAF with a well-matched shape but an incorrect mutation count; Or, in the opposite case, overemphasizing the mutation count term may produce accurate total mutations but a poorly fitted distribution.

To balance them, we first normalized the Wasserstein distance to a range between 0 and 1 to have it in the same scale as the second term of the equation. To do so, we divide the computed Wasserstein distance by its maximum possible value, obtained by comparing the observed VAF distribution to the most different possible one, ensuring the largest possible per-element differences.

Finally, we performed a random search on simulated* VAFs to determine the optimal value for the λ parameter. For each simulation, λ was sampled uniformly between 0 and 1, was executed with TAFI, and evaluated using the sum of absolute relative differences between the real values of the parameters (S and C) used to generate the simulations and the obtained ones. After evaluation, we found that $\lambda = 0.2$ provided the best trade-off between shape and scale accuracy (see Supplementary Figure 1a-d for validation).

** The process to generate simulated data is properly defined in Section 2.4. below.*

2.3.4. Model classification

The entire inference procedure described above is executed twice when running TAFI, once for each of the two evolutionary models, resulting in 20,000 simulations in total. To accelerate convergence for the exponential model, TAFI re-uses the best-fitting purity and C values obtained under the WF model as the starting points, based on the assumption that those parameters should be equivalent under both models, as they are independent of it.

After both optimization runs are complete, we compare the best results for each model based on their final loss values. The model that produces the prediction with the smallest loss —meaning the closest match to the observed data in terms of both VAF shape and mutation burden— is selected as the best solution for the sample's evolutionary history and the corresponding parameter values associated with it are reported.

2.4. Validation

To initially validate our results, we used simulations. We generated 200 synthetic VAF distributions by randomly sampling parameter values from predefined biologically plausible ranges and reconstructing by following the process explained in the Section 2.3.1 above. The parameters were sampled as follows: the evolutionary model was randomly chosen between WF and EXP; tumor purity ranged from 0.2 to 1.0; and the number of subclonal and clonal mutations ranged from 100 to 10^6 .

Each simulated sample was then processed using TAFI, and the predicted outputs were compared against the ground-truth values used in the simulation.

To assess the accuracy of model prediction, we constructed a confusion matrix comparing the predicted model classes (WF vs. EXP) against the true model classes. From this, we computed the following performance metrics: precision, recall, and F1-score (see equations 4-6 below) —which together reflect the model's predictive performance.

$$precision = \frac{TP}{TP + FP} \quad (4)$$

$$recall = \frac{TP}{TP + FN} \quad (5)$$

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall} \quad (6)$$

Here, TP (True Positives), FP (False Positives), and FN (False Negatives) were derived from the confusion matrix.

For the remaining parameters (purity, S, and C), we independently evaluated accuracy using the absolute relative error (equation 7) between the real parameters used to build the simulations and the predicted values using TAFI.

$$error = \frac{|real - predicted|}{real} \quad (7)$$

2.5. Downstream analyses

2.5.1. Differential gene expression analysis

We used the gene expression data (RNA-seq) available for our samples to investigate transcriptional differences between the two groups identified through TAFI classification. To do this, we performed a differential gene expression (DGE) analysis, a common bioinformatics approach used to identify differentially expressed genes (DEGs)—genes that show statistically significant differences in expression between two groups: the samples classified as WF, and the ones classified as EXP.

For this analysis, we used DESeq2 (version 1.48.0), a widely used R package for analyzing count-based RNA-seq data. DESeq2 models the data using a negative binomial distribution, then applies shrinkage estimation for dispersion and uses fold changes to improve the stability and interpretability of the results (Love, et al., 2014) [26]. Genes were considered differentially expressed based on their adjusted p-values ($FDR < 0.05$) and fold changes ($|\log_2(FC)| > 1$).

After identifying the DEGs, we classified them into upregulated (if they were significantly more expressed in the WF samples, compared to the EXP ones), or downregulated if it was the opposite case.

Then, we performed a gene set enrichment analysis (GSEA) on those sets to explore the biological functions in which those genes are involved. For this, we used the clusterProfiler R package (version 4.5.0), which enables statistical analysis and visualization of functional profiles of gene clusters. We focused our enrichment analysis on two main sources of functional annotation: The Gene Ontology (GO) Biological Process (BP) terms (The Gene Ontology Consortium, 2021) [27], which provide insights into the biological objectives and processes the DEGs are associated with; and the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (Kanehisa, et al., 2021) [28], which help to contextualize the genes within known signaling and metabolic pathways.

2.5.2. Survival analysis

Survival analysis is a statistical approach used to investigate the time until the occurrence of a specific event, usually death, in each sample. In cancer genomics, survival analysis is commonly employed to assess whether molecular or clinical factors are associated with differences in patient outcomes. We applied survival analysis to determine whether the evolutionary mode of a tumor, as predicted by TAFI (WF vs. EXP), is associated with significant differences in patient survival times from diagnosis.

We used the Cox proportional hazards model, a widely used method for modeling survival data, implemented via the `CoxPHFitter` function from the `lifelines` (version 0.30.0) Python package. The model estimates the effect of one or more covariates on the hazard function, which describes the instantaneous risk of the event (in this case, death) at any given time. The key assumption of the Cox model is that the effect of each covariate on the hazard is proportional over time, meaning that hazard ratios remain constant.

The primary metric we analyzed was Overall Survival (OS), defined as the time from diagnosis to death from any cause. Although other survival metrics exist, such as Progression-Free Interval (PFI) or Disease-Specific Survival, OS remains the most robust and widely used outcome metric. In our analysis, death was considered the event of interest, and patients who were still alive at the time of last follow-up were treated as right-censored observations.

To explore the association between evolutionary mode and survival, we included the predicted mode as a categorical covariate in the Cox model. We also incorporated age at diagnosis, tumor type, tumor stage, and gender of the patient as additional covariates to account for potential confounding effects. The output of the Cox model includes hazard ratios (HRs), which quantify the relative risk of the event occurring in one group compared to a reference group, and their associated p-value. A hazard ratio greater than 1 indicates increased risk, while a value less than 1 indicates reduced risk, relative to the reference group used for each comparison.

To complement the Cox regression results and provide a more intuitive, visual understanding of survival patterns, we also generated Kaplan-Meier survival curves stratified by evolutionary mode. Those curves estimate the probability of survival over time, allowing us to observe differences in survival distributions between groups without assuming any specific parametric form.

3. Results

3.1. Data exploration

From the initial 20,914 samples collected across all datasets, just 3403 samples passed all the filtering criteria defined by the sample selector. This corresponds to approximately 49% of the PCAWG samples, 31% of the HMF samples, and less than 1% of the MC3 cohort. The main reason for this sharp reduction, particularly in the MC3 data, is attributable to the 5th criterion of the filtering process: excluding samples with fewer than 1,000 mutations.

Figure 3 illustrates the distribution of tumor mutational burden (TMB), defined here as the total number of somatic mutations detected per sample. The TMB is shown on a $\log_{10}$ scale and colored by project to facilitate comparison. The plot reveals a clear difference between the MC3 cohort and the HMF/PCAWG datasets, showing that the MC3 samples are predominantly found in the low-mutation range. This discrepancy arises from the different sequencing strategies used in each project. Since the MC3 dataset is based on whole-exome sequencing (WES), it explores only the coding region of the genome. In contrast, both HMF and PCAWG are based on whole-genome sequencing (WGS), which covers the entire genome and therefore there are ≈ 50 times more sites that could have been mutated.

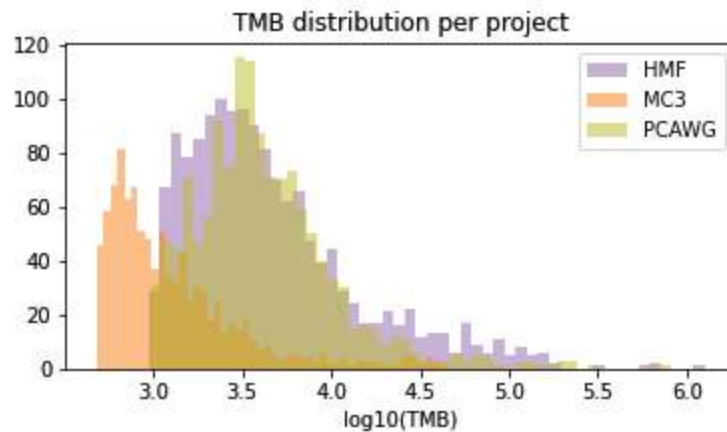


Figure 3. TMB distribution per project. Overlapping histograms showing the distribution of tumor mutational burden (in $\log_{10}$ scale) colored by project.

3.2. Predictions

3.2.1. On simulations

After running TAFI on 200 stochastic simulations, we evaluated its performance by comparing the predicted parameters to the true values used to generate each simulated sample. This allowed us to assess both the accuracy of evolutionary mode classification and the reliability of parameter inference.

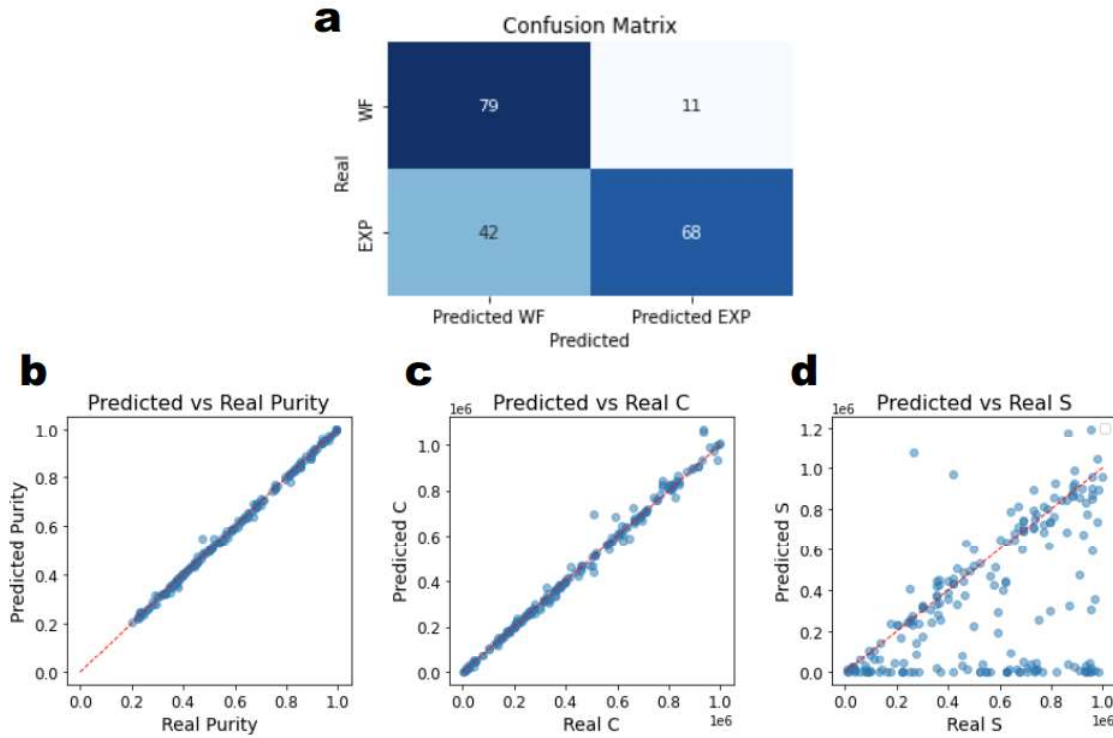


Figure 4. Real vs predicted parameters. (a) Confusion matrix showing the number of WF and EXP simulations, and its corresponding prediction obtained using TAFI on them. The bottom scatterplots show the real and predicted (b) purity , (c) C, and (d) S for the simulations, indicating a perfect prediction with a red dashed line in the diagonal.

Figure 4a shows how TAFI correctly classifies the majority of samples, demonstrating strong overall performance. However, the model shows a tendency to overpredict the WF category, even for samples that were generated under EXP dynamics, leading to an increased number of false positives for the WF class. This is reflected in the relatively lower precision for WF (0.65), compared to the higher precision for EXP (0.86). Despite this, the recall for WF remains high (0.88), indicating that most true WF cases are correctly identified.

When examining the continuous parameter inference, TAFI performs well for tumor purity and C . The mean absolute relative error for purity is just 1.22%, indicating highly accurate estimates across simulations (see Figure 1b). Similarly, for C , the mean error is just 3.6% (Figure 1c). These results show that TAFI reliably recovers both purity and clonal mutation burden, with small deviations from the true values. In contrast, the estimation of S is less accurate, reporting an average absolute relative error of 46.6% (Figure 1d). This deviation can be explained due to two primary factors. First, the number of observed subclonal mutations is highly sensitive to the stochastic simulated sampling process, introducing noise in the estimation. Second, S can vary substantially in scale between evolutionary models: in WF cases, the number of subclonal mutations tends to be moderate and better recovered, while in EXP simulations, S can be orders of magnitude larger. When such EXP samples are misclassified as WF, TAFI significantly underestimates the true subclonal burden. This explains the cluster of points in the figure where predicted S values are far lower than the true values.

Overall, these results on simulated data demonstrate that TAFI accurately infers purity and clonal structure, while subclonal inference is more challenging and highly sensitive to model classification.

3.2.2. On real data

Figure 5 below shows an example of a results plot generated after evaluating TAFI on a tumor sample (from PCAWG, in this case). The figure displays the variant allele frequency distribution for the best-fit prediction per model along with their associated parameters. In this particular case, by visual inspection alone, we can observe that the WF model best reproduces (i.e. overlaps) the observed VAF distribution, which is confirmed by the "distance" metric included in the plot (where distance refers to the loss function previously defined). This suggests that, for this particular tumor, the evolutionary dynamics are most likely described by the Wright-Fisher model —as evidenced by its smaller loss compared to the EXP model. For this reason, the inferred parameters include: a purity of 0.88, 2551 total subclonal mutations, and 2635 clonal mutations.

Notably, purity and C values (clonal mutation count) obtained for both models are highly consistent., which is expected since they are independent of the evolutionary model. In contrast, the subclonal mutation count (S) is model-dependent, scaling orders of magnitude for the EXP.

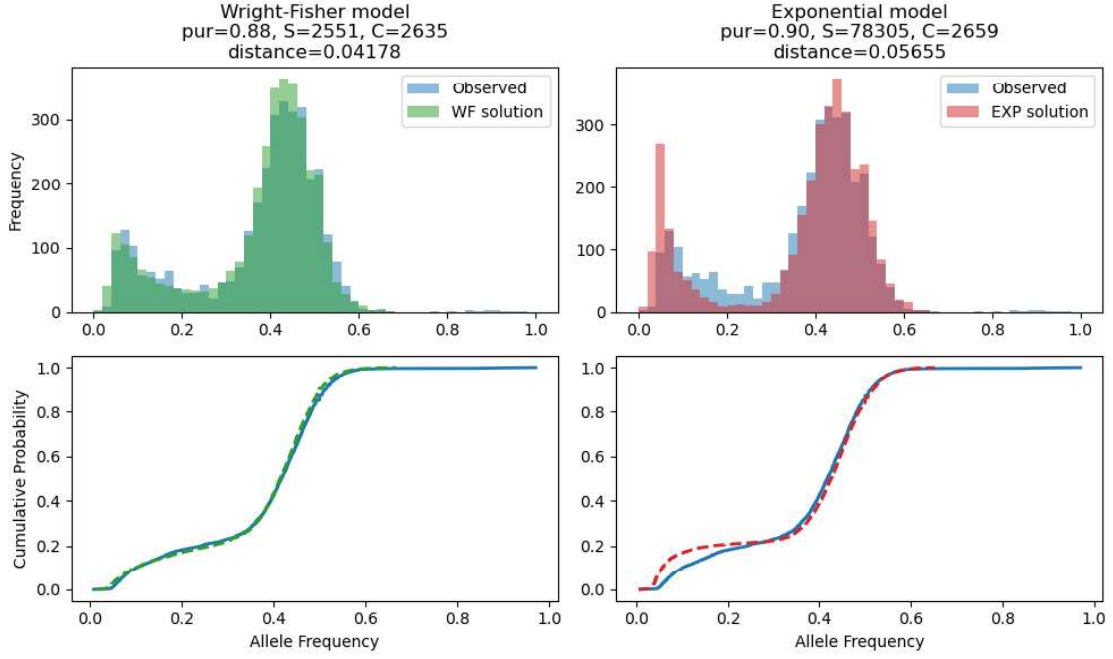


Figure 5. Example of a TAFI solution for a tumor sample. The plot consists of four panels arranged in a 2x2 grid, with each model displayed in a separate column and color-coded as follows: green for the Wright-Fisher (WF) solution, red for the exponential (EXP) solution, and blue for the observed (real) data. The top row shows a 50-bin histogram comparing the observed VAF distribution with the best-fit predictions from each model. The bottom row displays the corresponding cumulative density functions (CDFs) for the observed and predicted distributions, which are used to compute the Wasserstein distance being minimized.

The number of tumor samples per cohort assigned to each mode of growth is summarized in Table 2. In all datasets, a clear majority of samples were classified under the WF model, with relatively fewer tumors assigned to the EXP model. This trend is also illustrated in Figure 4a, which highlights a systematic bias in TAFI's classifications—suggesting that the algorithm more readily identifies WF-like evolutionary patterns, while EXP dynamics are less frequently assigned.

Notably, this imbalance is reduced in the HMF cohort, where the classification is closer to an even split between WF and EXP samples. This observation aligns with our initial hypothesis: since the HMF dataset predominantly consists of metastatic tumors, we expect these tumors to be more frequently characterized by rapid growth, consistent with the exponential model.

Project	WF	EXP	
MC3	610	45	655
PCAWG	1,070	231	1,301
HMF	778	669	1,447
	2,458	945	3,403

Table 2. TAFI classification. The table contains the number of samples classified as Wright-Fisher (*WF*) and exponential (*EXP*) for their tumor growth mode with TAFI. They are divided depending on which project they originally belong to.

Finally, we investigated the relationship between two of the key parameters inferred by TAFI: the number of clonal (*C*) and subclonal mutations (*S*). While the scatter plot in Figure 6 does not suggest a strong overall visual correlation between these two parameters, we quantified the association using the Pearson correlation coefficient.

Across all samples, we obtained a correlation of $r = 0.084$, with a significant p-value ($p < 1e^{-6}$), indicating a weak but significant positive correlation. To explore whether this relationship differed by evolutionary model, we computed separate correlations for samples classified as WF and EXP. Interestingly, within the WF group (green points in the plot), we observed a positive correlation of $r = 0.398$, with an extremely significant p-value ($p < 1e^{-93}$). In contrast, for tumors classified as EXP, the correlation was essentially negligible ($r = 0.024$, $p = 0.45$), indicating no significant linear relationship between *S* under exponential growth dynamics.

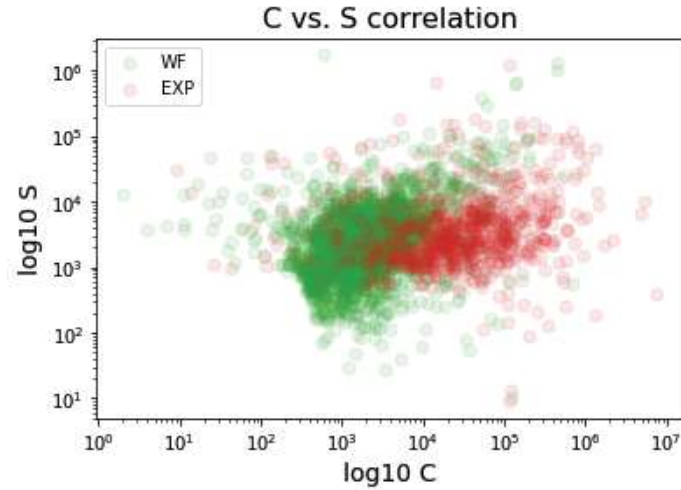


Figure 6. Correlation between S and C. Scatterplot showing the relationship between the predicted values of subclonal (S) and clonal (C) mutations, plotted in logarithmic space. Points are colored according to the evolutionary model assigned by TAFI.

3.3. Downstream analyses

3.3.1. Differential gene expression

RNA-seq data is inherently high-dimensional, containing expression measurements for thousands of genes per sample, generating data frames with millions of fields. For this reason, dimensionality reduction techniques, such as Principal Component Analysis (PCA), are commonly applied as a first step to do data exploration. These methods aim to project the data into a lower-dimensional space that captures the largest sources of variance in the dataset, making it easier to identify major trends or sample groupings.

This approach was applied to the HMF dataset, and Figure 7 displays the result of plotting the first two principal components. To better understand what factors explain the variation, we colored the samples according to two variables: the predicted evolutionary model (WF vs. EXP) and the tumor's primary location.

The figure reveals that the dominant source of variation is not the inferred evolutionary model, but rather the tumor's primary location, suggesting that the tissue of origin must be accounted for. For this reason, we included tumor location as a covariate in our differential expression design formula, modeling gene expression as $\sim \text{type} + \text{mode}$, where “type” refers to tumor location and “mode” corresponds to the evolutionary classification inferred by TAFI.

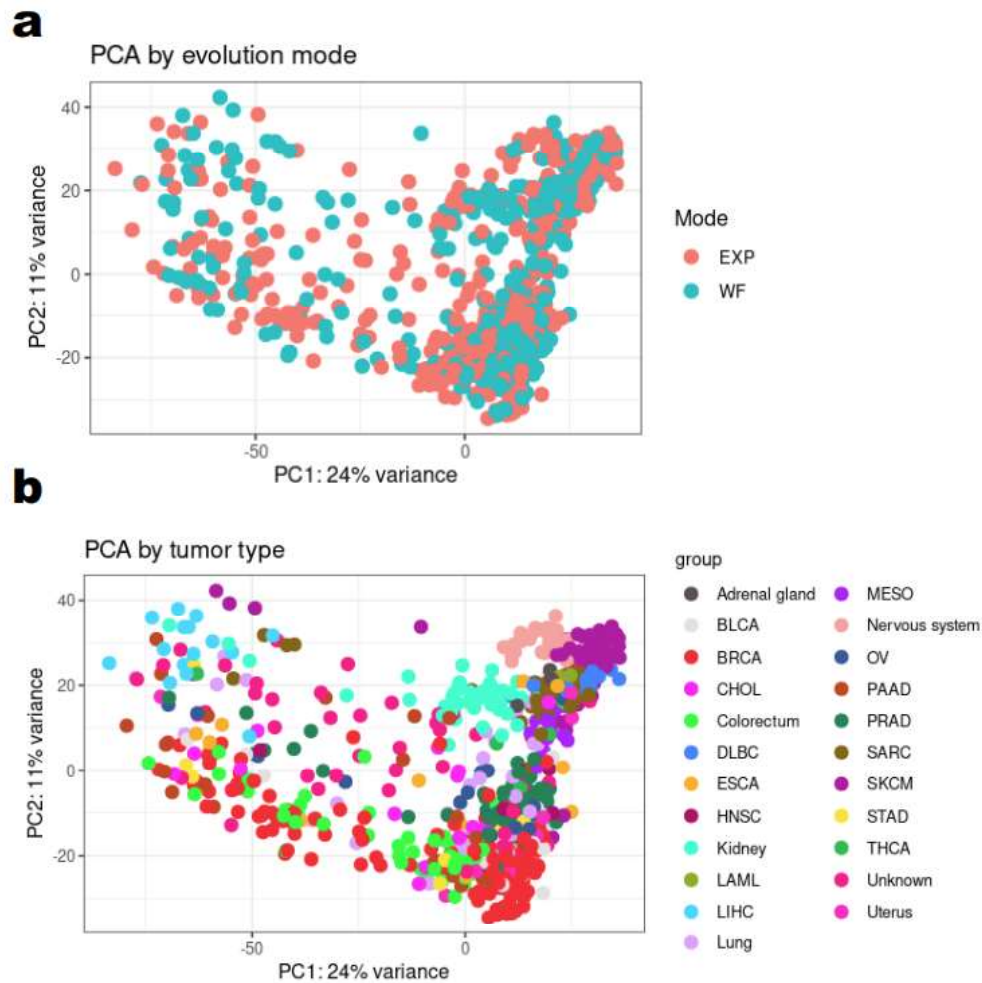


Figure 7. Principal component analysis (PCA) of gene expression data. The first two principal components (PC1 and PC2) are shown, colored by (a) mode and (b) tumor location.

After performing differential expression analysis using DESeq2, we identified genes that are significantly differentially expressed depending on the mode of evolution, using the WF group as the baseline condition.

Based on adjusted p-values and fold-changes, we identified a total of 110 differentially expressed genes (DEGs). Among them, 63 genes were upregulated, meaning they were more highly expressed in WF-classified tumors compared to EXP-classified tumors. In contrast, 47 genes were downregulated, showing lower expression levels in WF samples relative to the EXP group (see Supplementary Figure 3).

	Code	Description	Adjusted p-value
WF > EXP	GO:0003012	Muscle system process	< 0.05
	GO:0006936	Muscle contraction	< 0.05
	GO:0042692	Heart contraction	< 0.05
	KEGG:hsa04820	Cardiac muscle contraction	< 0.01
	KEGG:hsa05410	Hypertrophic cardiomyopathy	< 0.01
	KEGG:hsa05414	Dilated cardiomyopathy	< 0.01
EXP > WF	GO:0009913	Epidermal cell differentiation	< 0.05
	GO:0043588	Skin development	< 0.05
	KEGG:hsa04082	Neuroactive ligand signaling	< 0.05

Table 3. Enriched functions of the DEGs. Summary of top-ranked significantly enriched GO biological processes and KEGG pathways in upregulated (WF > EXP) and downregulated (EXP > WF) genes.

We performed functional enrichment analysis on the upregulated and downregulated gene sets to know the biological implications of the DEGs. A summary of the top-ranked significant biological processes and pathways is shown in Table 3.

The upregulated genes were significantly enriched in muscle-related biological processes, including muscle system process, muscle contraction, and heart contraction, while enriched pathways included cardiac muscle contraction and different kinds of cardiomyopathies.

In contrast, the downregulated gene set showed a weak and limited enrichment in skin-related processes, such as epidermal cell differentiation and skin development, as well as the KEGG pathway neuroactive ligand-receptor interaction.

3.3.2. Survival analysis

Table 4 summarizes the results of the Cox proportional hazards model to evaluate the impact of several covariates on overall survival times on the PCAWG samples. Below, we summarize the main findings.

- Age: Age at diagnosis was significantly associated with overall survival. Specifically, each additional year of age increased the hazard of death by approximately 2% (HR = 1.02). Although this effect is modest, it is statistically significant and consistent with the fact that older patients tend to have poorer outcomes in terms of therapy response.
- Gender: Gender was also a significant predictor of survival where males exhibited a higher risk of death compared to females.
- Tumor Stage: Clinical stage at diagnosis showed a positive association with hazard. Patients with later-stage tumors had higher risk of death, which aligns with clinical expectations, as advanced-stage cancers are generally more aggressive and less responsive to treatment.
- Tumor Type: Certain tumor types were significantly associated with reduced survival. In particular, patients with Cervical Squamous Cell Carcinoma (CESC), Glioblastoma Multiforme (GBM), Brain Lower Grade Glioma (LGG), and Sarcoma (SARC) showed significantly higher hazards compared to the reference. These findings reflect known differences in prognosis across cancer types (Siegel et al., 2024)[29]. Other tumor types included in the analysis did not show statistically significant associations with survival in this model.
- Evolutionary Mode: Finally, we examined the impact of the tumor's evolutionary model classified by TAFI, on survival. Tumors classified as EXP had a 25% higher hazard of death compared to those classified as WF (HR = 1.25). However, this difference was not statistically significant ($p = 0.40$), indicating that we cannot confidently conclude an association between the predicted evolutionary mode and patient survival based on this test.

Overall, the Cox model identifies several well-established prognostic factors —like age, gender, stage, and tumor type— as significantly associated with survival. However, the predicted evolutionary mode does not show a significant effect, although the median survival time for the WF-samples almost doubles the one for the patients classified as EXP (Figure 8).

Factor	HR	p-value
Age	1.02	0.03
Gender	1.46	0.05
Stage	1.20	<0.005
Mode	1.25	0.40
type_CESC	23.96	<0.005
type_GBM	47.46	<0.005
type_LGG	12.51	<0.005
type_SARC	7.30	0.02
...	...	...

Table 4. Cox proportional hazard significant results. Hazard ratios and their associated p-value for age, gender, tumor stage, evolutionary mode and cancer type (only significant types are shown. Supplementary Table 2 stores the whole information).

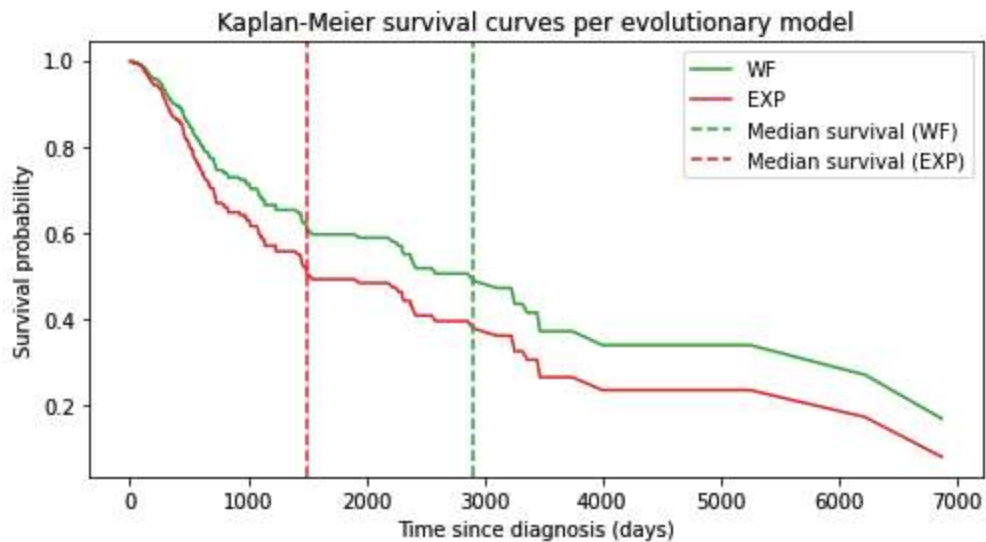


Figure 8. Kaplan-Meier curves per mode. Survival curves showing the evolution in the survival probability over time since the first diagnosis of the patient. Dashed vertical lines indicate the median survival time for each group: 1500 days for the EXP group, and 2907 for the WF one.

3.4. TAFI vs. MOBSTER

MOBSTER can be considered the state-of-the-art tool for inferring the evolutionary mode of tumors. A key parameter in MOBSTER's inference is the “shape” parameter, which modulates the decay of this Pareto tail. In the context of TAFI, this parameter can be viewed as conceptually translated as 1 corresponds for the WF model, or 2 for the EXP one.

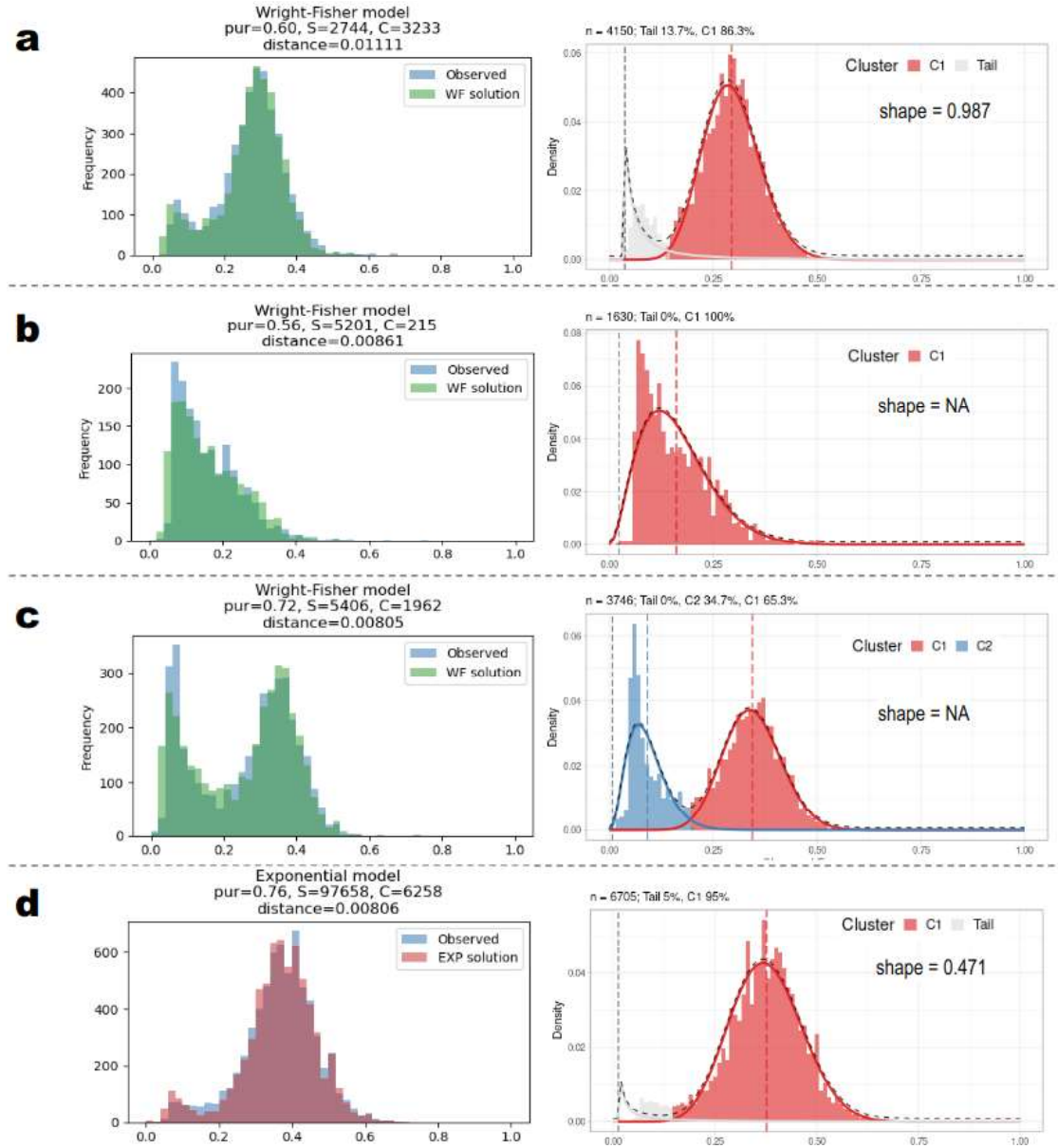


Figure 9. TAFI and MOBSTER fit comparison. Four representative VAF fits from real tumor samples using TAFI (left) and MOBSTER (right), illustrating cases of agreement, clonal misclassification, subclone misinterpretation, and shape underestimation.

To compare the performance of the two methods, we ran MOBSTER on the same set of 1,301 PCAWG samples that were previously analyzed with TAFI. Although MOBSTER successfully produced fits for the majority of them, only 62 samples reported a valid shape value. This overlap highlights a limitation of MOBSTER in front of TAFI.

Figure 9 illustrates representative examples of how TAFI and MOBSTER fit real VAF distributions across four typical scenarios.

- 9a. Both TAFI and MOBSTER agree, providing consistent interpretations of the data and similar estimates of the underlying shape parameter.
- 9b. MOBSTER fails to detect a subclonal component and instead fits the entire VAF distribution as a single clonal population, while TAFI correctly identifies a superposition of clonal and subclonal peaks.
- 9c. MOBSTER mistakenly interprets the subclonal peak as a segregating subclone, resulting in a misclassification.
- 9d. MOBSTER underestimates the subclonal peak, leading to a lower inferred shape parameter.

These differences in fit behavior help to explain the patterns observed in Figure 10, where we compare the shape values inferred by MOBSTER to the binary evolutionary classifications produced by TAFI. While the average shape estimated for WF-classified tumors is relatively consistent and centered around 1, as expected, the shape values for EXP-classified samples are more dispersed and underestimated. This bias likely reflects MOBSTER's tendency to miss or misinterpret subclonal peaks in fast-evolving tumors, particularly in the presence of complex VAF structures.

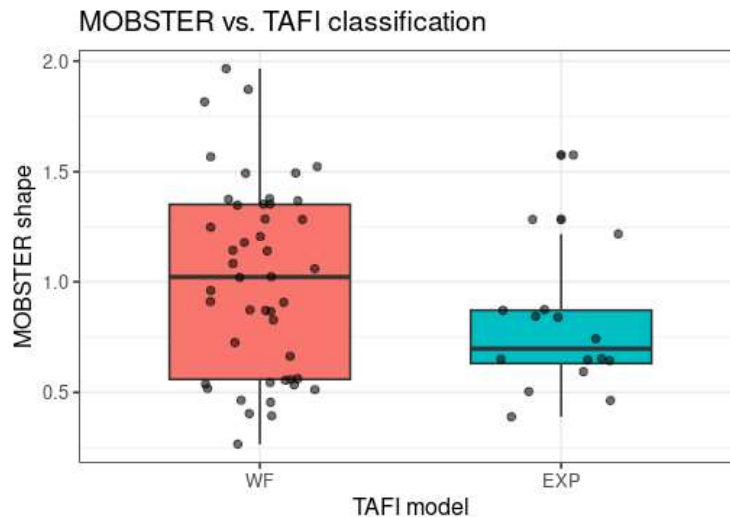


Figure 10. Comparison of MOBSTER and TAFI evolutionary classification. Boxplots show the distribution of MOBSTER-inferred shape values stratified by TAFI-predicted model class (WF = 1, EXP = 2).

4. Discussion

This thesis highlights the potential of variant allele frequencies as a window into the evolutionary history of tumors. By studying the distribution of VAFs derived from bulk sequencing data, we can infer not only the current clonal architecture of a tumor but also the evolutionary processes that have shaped it, offering relevant clinical and prognostic information. Understanding these dynamics is essential for potentially improving cancer diagnosis, stratifying patients, and guiding therapeutic strategies.

We introduced TAFI (Tumor Allele Frequency Interpreter), a novel simulation-based framework for inferring tumor evolutionary dynamics from VAF data. TAFI provides an approach to classify tumors into two canonical evolutionary modes: Wright-Fisher (WF), inspired by the constant population size model from population genetics; and Exponential (EXP), representing a tumor growing exponentially. In addition to classification, TAFI estimates three key parameters: tumor purity, the number of clonal mutations (C), and the true number of subclonal mutations (S). Unlike previous methods, TAFI explicitly models the full stochasticity of the sequencing process by simulating VAF distributions under user-defined parameters, which allows the framework to reconstruct the generative processes behind observed data more accurately.

TAFI demonstrates strong performance on synthetic data, accurately recovering tumor purity and clonal mutation counts with low error rates. However, the inference of subclonal mutation counts is more variable, reflecting the difficulty of estimating subclonal diversity.

When applied to large-scale cancer datasets—including PCAWG, HMF, and MC3—TAFI classifies the majority of tumors as WF. Interestingly, this distribution shifts in the HMF cohort, which consists primarily of metastatic tumors. In this dataset, TAFI assigns a nearly balanced number of samples to each evolutionary mode, supporting the initial hypothesis that advanced, metastatic tumors are more likely to follow exponential growth dynamics. This finding reinforces the biological relevance of the model and suggests that evolutionary mode may reflect tumor aggressiveness and progression stage.

A particularly relevant result emerged from the correlation analysis between clonal and subclonal mutation counts. While no correlation was found in EXP-classified tumors, we observed a significant positive correlation in WF tumors. This suggests that in constantly evolving tumors, mutation rate contributes proportionally to both clonal and subclonal burden, while in exponential tumors there might be a factor that perturbs this correlation.

To assess whether the evolutionary classification reflects deeper biological differences, we conducted first differential gene expression analysis. Tumors classified as WF showed significant enrichment in muscle and cardiac-related genes, including pathways linked to cardiac muscle contraction and cardiomyopathy, while EXP-classified tumors were enriched in epidermal development and neuroactive ligand-receptor signaling. From those findings we highlight the appearance of cardiomyopathy, a heart disease, which has been associated with chemotherapy treatments in the literature (Shakir & Rasul, 2009) [31] before.

We also examined whether the evolutionary mode inferred by TAFI has prognostic value. Using overall survival (OS) as the outcome, we found that traditional clinical variables (age, gender, tumor stage, and tumor type) were significantly associated with patient outcomes. Evolutionary mode, however, did not reach statistical significance in this dataset. Nevertheless, we observed a trend toward longer median survival in WF-classified tumors compared to EXP. Although not conclusive, this result suggests a potential prognostic signal that may become more evident in larger datasets.

Compared to existing methods like MOBSTER (the state-of-the-art), TAFI demonstrates a better ability to fit subclonal architectures, especially in EXP-classified tumors. This leads to a more accurate estimation of the shape parameter (α), which reveals the evolutionary mode of the tumor. By focusing on the full generative process and incorporating stochastic modeling, TAFI represents a meaningful advancement in the quantitative study of cancer evolution.

5. Conclusions

5.1. Limitations

While we have demonstrated TAFI performance, there are several limitations to its current implementation that affect its generalizability and performance.

The first limitation is the number of samples for which TAFI is applicable. Despite collecting a large dataset, the stringent filtering criteria used in the current version of TAFI's sample selector resulted in a huge reduction in usable samples, where only 6.3% of WES and 27.7% of WGS samples were retained. This reduction was necessary to ensure clean and interpretable VAF distributions for initial validation, but it significantly limited the size of the final cohort and reduced the statistical power of downstream analyses.

A related limitation lies in the assumptions made by the current version of the algorithm. TAFI is designed to model at most two distinct peaks in the VAF distribution, representing a clonal and a subclonal populations, and excludes samples that contain segregating subclones or more complex architectures. These cases, however, represent a substantial portion of real tumor samples.

Another notable limitation is the model's bias toward classifying tumors as the WF model. TAFI tends to overpredict the WF category, even in cases where the sample follows an EXP model. This misclassification not only impacts the evolutionary interpretation but also leads to systematic errors in parameter estimation, particularly for the number of subclonal mutations (S), which is dependent on the assumed mode of evolution.

Finally, compared to tools like MOBSTER, TAFI currently assigns tumors to only two discrete values of the Pareto shape parameter (α): 1 for WF and 2 for EXP. While this binary approach simplifies classification and aligns with population genetics theory, it limits the model's ability to capture intermediate or mixed evolutionary dynamics, which are common in many tumors. As a result, samples are forced into one of the categories, potentially oversimplifying the processes.

5.2. Future work

The next phase of this project will focus on the development and release of a second version of TAFI, aimed at addressing the key limitations identified in the current implementation.

A primary objective is to increase the flexibility of the sample selection criteria, which in this first version were deliberately stringent in order to optimize performance and ensure clean input data. In particular, we plan to lower the minimum mutation threshold required for sample inclusion. Although this adjustment may introduce slightly noisier data, the threshold will be carefully chosen to ensure that the algorithm can still perform robustly. This modification will allow a significantly larger fraction of real-world samples to be processed by TAFI, especially those sequenced using a whole-exome sequencing approach.

Another major extension will be the incorporation of models for segregating subclonal populations. Currently, samples exhibiting more than two distinct peaks in their VAF distributions are excluded, as the algorithm assumes a simple clonal-subclonal architecture. However, many tumors exhibit more complex evolutionary patterns involving multiple subclonal lineages. Expanding TAFI's framework to account for these multi-subclonal architectures would increase the number of analyzable samples and improve its applicability to advanced or heterogeneous tumors.

In addition, we plan to move from the current binary classification of the evolutionary mode to a more flexible model capable of inferring α as a continuous parameter. This enhancement would allow TAFI to capture intermediate evolutionary regimes that do not conform neatly to the extremes of neutral or strong selection. Such an upgrade would improve TAFI evolutionary classifications, particularly in tumors exhibiting transitional dynamics.

Following these methodological improvements, we intend to re-run TAFI on the original dataset, but this time including 17,418 additional whole-genome cancer samples from the 100,000 Genomes Project, conducted by the Genomics England consortium [32]. This will significantly increase the statistical power, which is a current limitation.

Finally, we will also expand the scope of downstream analyses. In particular, we plan to perform a differential selection coefficient analysis using CBaSE (Weghorn & Sunyaev, 2017) [33], another tool developed in our lab. This analysis aims to identify genes under differential selective pressure across different evolutionary modes that could potentially serve as mode-specific driver genes.

5.3. Ethical and diversity aspects

This study relies on large-scale human genomic data, which follows privacy and consent protocols. All datasets (PCAWG, MC3, HMF) were under controlled access and data-use agreements that include patient informed consent. No individual VCFs or clinical metadata were exposed in our analyses; all work was conducted on de-identified data in secure environments. We ensured compliance with GDPR for European cohorts.

The datasets used in this study include patients covering multiple cancer types and geographic regions. However, as is common in large-scale genomic dataset, individuals of European ancestry remain overrepresented. While important to consider, it should not represent a limitation for our method. TAFI does not incorporate demographic features, and is applicable on any patient sample, regardless of ethnicity, gender, or age.

Code Availability

TAFI and all the scripts used for this project are available in the following GitHub repository: <https://github.com/raulsanr/TAFI>. Please note that at submission date (06/06/2025) the repository is temporarily private, but will be made publicly accessible upon publication of the corresponding paper.

In the meantime, you can request access to the scripts by contacting: raul.sanz@crg.eu.

References

1. Greaves, M., & Maley, C. C. (2012). Clonal evolution in cancer. *Nature*, 481(7381), 306-313. <https://doi.org/10.1038/nature10762>
2. Beerenwinkel, N., Schwarz, R. F., Gerstung, M., & Markowetz, F. (2015). Cancer evolution: mathematical models and computational inference. *Systematic Biology*, 64(1), e1-e25. <https://doi.org/10.1093/sysbio/syu081>
3. Martincorena, I., & Campbell, P. J. (2015). Somatic mutation in cancer and normal cells. *Science*, 349(6255), 1483-1489. <https://doi.org/10.1126/science.aab4082>
4. Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., & Kinzler, K. W. (2013). Cancer genome landscapes. *Science*, 339(6127), 1546-1558. <https://doi.org/10.1126/science.1235122>
5. McFarland, C. D., Korolev, K. S., & Mirny, L. A. (2013). Impact of deleterious passenger mutations on cancer progression. *Proceedings of the National Academy of Sciences*, 110(8), 2910-2915. <https://doi.org/10.1073/pnas.1213968110>
6. Bozic, I., Antal, T., Ohtsuki, H., Carter, H., Kim, D., Chen, S., ... Nowak, M. A. (2010). Accumulation of driver and passenger mutations during tumor progression. *Proceedings of the National Academy of Sciences*, 107(43), 18545-18550. <https://doi.org/10.1073/pnas.1010978107>
7. Fisher, R. A. (1930). *The genetical theory of natural selection*. Clarendon Press. <https://doi.org/10.5962/bhl.title.27468>
8. Wright, S. (1931). Evolution in Mendelian populations. *Genetics*, 16(2), 97-159. <https://doi.org/10.1093/genetics/16.2.97>
9. Williams, M. J., Werner, B., Heide, T., Curtis, C., Barnes, C. P., & Sottoriva, A. (2016). Quantification of subclonal selection in cancer from bulk sequencing data. *Nature Genetics*, 48(7), 895-902. <https://doi.org/10.1038/ng.3604>
10. Sottoriva, A., Kang, H., Ma, Z., Graham, T. A., Salomon, M. P., Zhao, J., ... Curtis, C. (2015). A Big Bang model of human colorectal tumor growth. *Nature Genetics*, 47(3), 209-216. <https://doi.org/10.1038/ng.3214>
11. Yates, L. R., & Campbell, P. J. (2012). Evolution of the cancer genome. *Nature Reviews Genetics*, 13(11), 795-806. <https://doi.org/10.1038/nrg3317>

12. International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. *Nature* 431, 931-945. <https://doi.org/10.1038/nature03001>
13. Metzker, M. L. (2010). Sequencing technologies —the next generation. *Nature Reviews Genetics*, 11(1), 31-46. <https://doi.org/10.1038/nrg2626>
14. Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, 26(10), 1135-1145. <https://doi.org/10.1038/nrg2626>
15. van Dijk, E. L., Auger, H., Jaszczyszyn, Y., & Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in genetics : TIG*, 30(9), 418-426. <https://doi.org/10.1016/j.tig.2014.07.001>
16. Choi, M., Scholl, U. I., Ji, W., Liu, T., Tikhonova, I. R., Zumbo, P., Nayir, A., Bakkaloğlu, A., Ozen, S., Sanjad, S., Nelson-Williams, C., Farhi, A., Mane, S., & Lifton, R. P. (2009). Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, 106(45), 19096-19101. <https://doi.org/10.1073/pnas.0910672106>
17. Schneider, V. A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H. C., Kitts, P. A., Murphy, T. D., Pruitt, K. D., Thibaud-Nissen, F., Albracht, D., Fulton, R. S., Kremitzki, M., Magrini, V., Markovic, C., McGrath, S., Steinberg, K. M., Auger, K., Chow, W., Collins, J., Harden, G., ... Church, D. M. (2017). Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome research*, 27(5), 849-864. <https://doi.org/10.1101/gr.213611.116>
18. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 10, 57-63 (2009). <https://doi.org/10.1038/nrg2484>
19. Graham, T. A., & Sottoriva, A. (2017). Measuring cancer evolution from the genome. *The Journal of pathology*, 241(2), 183-191. <https://doi.org/10.1002/path.4821>
20. McGranahan, N., & Swanton, C. (2017). Clonal heterogeneity and tumor evolution: Past, present, and the future. *Cell*, 168(4), 613-628. <https://doi.org/10.1016/j.cell.2017.01.018>
21. Nik-Zainal, S., Van Loo, P., Wedge, D. C., Alexandrov, L. B., Greenman, C. D., Lau, K. W., ... & Campbell, P. J. (2012). The life history of 21 breast cancers. *Cell*, 149(5), 994-1007. <https://doi.org/10.1016/j.cell.2012.04.023>
22. Caravagna, G., Heide, T., Williams, M. J., Zapata, L., Nichol, D., Chkhaidze, K., Cross, W., Cresswell, G. D., Werner, B., Acar, A., Chesler, L., Barnes, C. P., Sanguinetti, G., Graham, T. A., & Sottoriva, A. (2020). Subclonal reconstruction of tumors by using machine learning and population genetics. *Nature genetics*, 52(9), 898-907. <https://doi.org/10.1038/s41588-020-0675-5>
23. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (2020). Pan-cancer analysis of whole genomes. *Nature* 578, 82-93 <https://doi.org/10.1038/s41586-020-1969-6>

24. Ellrott, K., Bailey, M. H., Saksena, G., Covington, K. R., Kandoth, C., Stewart, C., Hess, J., Ma, S., Chiotti, K. E., McLellan, M., Sofia, H. J., Hutter, C., Getz, G., Wheeler, D., Ding, L., MC3 Working Group, & Cancer Genome Atlas Research Network (2018). Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *Cell systems*, 6(3), 271-281.e7. <https://doi.org/10.1016/j.cels.2018.03.002>
25. Hartwig Medical Foundation. (n.d.). Database. Retrieved from <https://www.hartwigmedicalfoundation.nl/en/data/database/>
26. Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15, 550. <https://doi.org/10.1186/s13059-014-0550-8>
27. The Gene Ontology Consortium. (2021). The Gene Ontology resource: Enriching a Gold mine. *Nucleic Acids Research*, 49(D1), D325-D334. <https://doi.org/10.1093/nar/gkaa1113>
28. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2021). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, 49(D1), D545-D551. <https://doi.org/10.1093/nar/gkaa970>
29. Siegel, R. L., Giaquinto, A. N., & Jemal, A. (2024). Cancer statistics, 2024. *CA: A Cancer Journal for Clinicians*, 74(1), 12–49. <https://doi.org/10.3322/caac.21820>
30. Shakir, D. K., & Rasul, K. I. (2009). Chemotherapy induced cardiomyopathy: pathogenesis, monitoring and management. *Journal of clinical medicine research*, 1(1), 8-12. <https://doi.org/10.4021/jocmr2009.02.1225>
31. Genomics England. (n.d.). 100,000 Genomes Project. Retrieved from <https://www.genomicsengland.co.uk/initiatives/100000-genomes-project>
32. Weghorn, D., Sunyaev, S. Bayesian inference of negative and positive selection in human cancers. *Nat Genet* 49, 1785-1788 (2017). <https://doi.org/10.1038/ng.3987>