Understanding the evolution of tumours using hybrid deep generative models

Anonymous Authors[1]

Abstract

Understanding both the population or subclonal structure and evolutionary forces that drive tumour evolution has important clinical implications for patients. However, deconvoluting subclonal structure and performing evolutionary parameter inference have largely been treated as two independent or step-wise tasks. Here, we show that combining stochastic simulations with hybrid deep generative models enables *joint* inference of subclonal structure and evolutionary parameter estimates. Ultimately, by jointly learning these two tasks, we show that our proposed approach leads to improve performance across a multitude of cancer evolution tasks including, but not limited to, detecting subclones, quantifying subclone frequency, and estimating mutation rate. As an additional benefit, we also show that hybrid deep generative models also provide substantial reductions in inference time relative to existing methods.

1. Introduction

1.1. Cancer evolution inference

The observation that evolutionary processes such as mutation, genetic drift, and selection can shape the heterogeneity, adaptability, and growth trajectory of tumour cell populations has made understanding the evolutionary or subclonal dynamics in patient tumours one of the major goals of cancer genomics (Black & McGranahan, 2021). Notably, quantifying and classifying evolutionary dynamics in biopsied and "bulk" DNA sequenced tumours has been shown to help stratify patient risk, identify critical subpopulations of cells, predict future clinical progression, and uncover important information on the underlying genetic determinants of tumour development (Fittall & Van Loo, 2019). And while much attention has focused on new single cell genomic applications, clinically, the majority of tumor biopsies will continue to be sequenced in "bulk" (Berger & Mardis, 2018). Thus, emphasizing the necessity for continued development of evolutionary inference methods that support this data type and, by proxy, future personalized genomics protocols.

1.2. Detecting subclones and quantifying selection

In general, cancer evolutionary inference from noisy 'bulk DNA sequenced tumours has been split across two fundamental tasks. The first task defined as subclonal clustering is focused on deconvoluting subpopulations of cells with similar characteristics such as mutational haplotypes (Dentro et al., 2017). The second task has been to use population genetic models of tumour evolution to infer parameters, such as mutation rate or subclone fitness, that are consistent with the observed data (Bozic & Wu, 2020). Together, subclonal clustering provides insight into the underlying cellular composition and genetic architecture of the tumour while evolutionary parameter inference helps inform on quantitative or qualitative properties related to the historical development and future growth of the tumour.

In the first task of subclonal clustering, the guiding principle is that when tumour biopsies are DNA sequenced each subpopulation of cells, i.e. subclones, will carry a unique set of mutations that will be 'sampled' in proportion to their cellular fraction within the tumour. Practically, in the context of single nucleotide or insertion-deletion mutations, this means that the sequencing read count or variant allele (mutation) frequency (VAF) distributions should be a mixture representing different subpopulations of cells. Evidently, subclonal clustering has largely been framed as a mixture model problem which has inspired the development of a variety of nonparametric or parametric methods using beta, binomial, or beta-binomial mixtures to describe the observed VAF or read count distributions in sequenced tumour biopsies (Nik-Zainal et al., 2012; Miller et al., 2014; Gillis & Roth, 2020). However, recent seminal work has shown that early subclonal clustering methods systematically overestimate the number of true subclonal populations unless the underlying evolutionary process is taken into account (Caravagna et al., 2020); namely that low VAF mutations do not represent a single subclonal population but rather a polyphyletic cluster or neutral power law 'tail' that is attributed to new mutations accruing in all dividing cells in the growing tumour.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

In the second task of evolutionary parameter inference, the goal has been to use the theoretical connection between observed mutation frequencies (VAFs) and underlying models of tumour growth to differentiate between competing models of evolution and to assign quantitative parameters based on these models to the observed data. Existing approaches to estimate evolutionary parameters, such as mutation rate, subclone fitness, and subclone emergence, time have primarily utilized classic likelihood-free methods such as approximate Bayesian computation (ABC) (Williams et al., 2018), parametric bootstrap methods that connect re-sampled mutation counts from mixture model fits to analytical evolutionary theory (Caravagna et al., 2020), or synthetic supervised learning which combines simulations and discriminative neural networks (Ouellette & Awadalla, 2022)

1.3. New strategies for understanding tumour evolution

A downside of existing approaches is that both of the above tasks are treated and inferred independently and, therefore, estimated in a stepwise manner, i.e. detecting mixtures before parameters or vice versa. However, theoretical work across evolutionary systems has shown that understanding subpopulation structure can inform on evolutionary parameters, making simultaneous estimation of both tasks important.

Here, we present *preliminary* work that takes advantage of recent advances in neural simulation-based inference and deep generative modeling to *jointly* infer subclonal mixtures and evolutionary parameters using stochastic simulations of tumour evolution and end-to-end differentiable hybrid variational autoencoders (HVAE). We provide *early experiments* showing improved accuracy and performance across multiple cancer evolution tasks, relative to existing methods, and provide a brief commentary on future improvements.

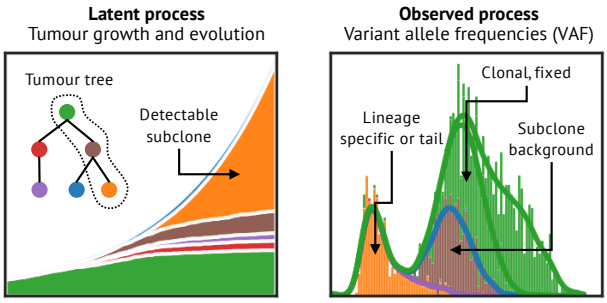

Figure 1. The connection between the latent process of tumour growth and evolution (left) and the observable process captured in the distribution of variant allele frequencies (right).

2. Methods

2.1. Setup

The frequency of observed mutations in bulk sequenced tumour biopsies encode information about the underlying evolutionary dynamics (Figure 1). Therefore, to perform neural simulation-based inference, we require a *simulator* $f(\mathbf{x}|\theta, \mathcal{M})$ that generates synthetic tumour sequencing data, and a *prior* that specifies our assumptions about the set of plausible parameters and models that define the evolutionary process $p(\theta, \mathcal{M})$.

An important nuance in notation is that we allow for θ to also specify latent information generated and collected during simulation, in addition to user-specified simulation parameters.

By generating millions of realizations from the joint distribution $f(\mathbf{x}|\theta, \mathcal{M})p(\theta, \mathcal{M})$ and building a synthetic dataset $\mathcal{D} = \{(x_i, \theta_i, \mathcal{M}_i)\}_{1:D}$, we can then train a HVAE to:

1. Decompose the observed VAF distribution[1] $\in \mathbb{R}^{1\times b}$ into a matrix of component VAF distributions $\in \mathbb{R}^{k\times b}$ where b represents the number of bins used to generate a VAF distribution and k indicates the number of mixtures representing neutral, subclonal, and clonal components
2. Perform model selection $p(\mathcal{M}|x)$ and provide approximate posterior estimates for evolutionary parameters of interest $p(\theta|x)$

Here, the HVAE is trained to perform *amortized inference*. This is due to time and computation expensive simulations that are not suited for sequential methods and as additional latent information generated during simulation is also recorded for downstream inference.

2.2. Simulator — $x_i \sim f(\mathbf{x}|\theta_i, \mathcal{M}_i)$

Previous computational and experimental work has shown that individual tumours generally follow an exponential growth trajectory. Therefore, we implemented a simulation framework based on a stochastic branching process model of exponential tumour growth and evolution, similar to previous work (Waclaw et al., 2015; Williams et al., 2018; Ouellette & Awadalla, 2022). A complete description of the simulator is provided in the Appendix A.1.1.

Overall, a single simulation returns a tuple that includes a synthetic VAF distribution $x_i \in \mathbb{R}^{1\times b}$ that represents the observable mutation data measurable from sequenced tumour biopsies, a set of synthetic VAF distributions representing

[1] A VAF distribution is simply a histogram of the variant allele frequencies and provides a way to standardize the input feature size to neural networks

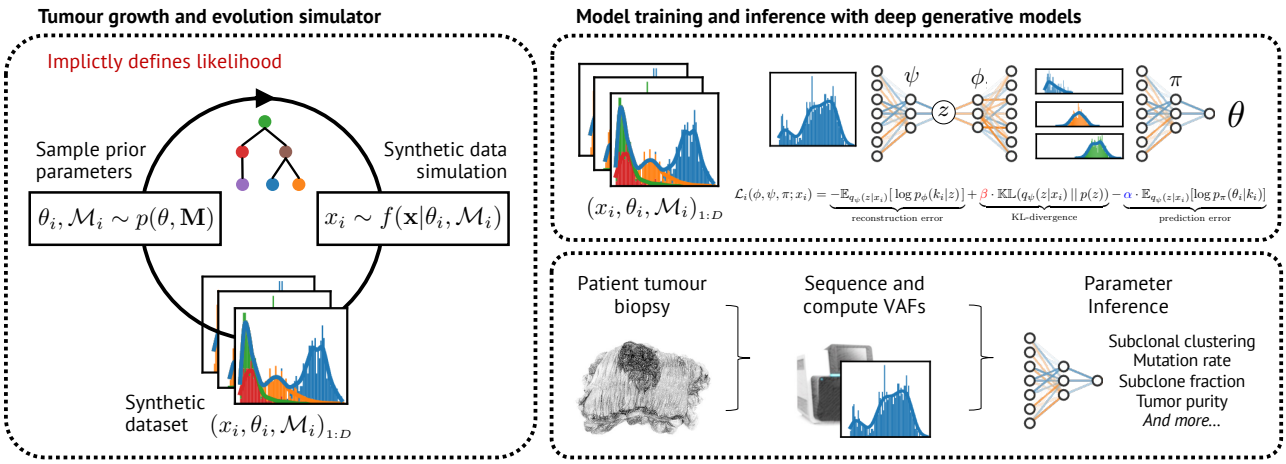

Figure 2. A stochastic generative process or simulator implicitly defines the likelihood of tumour sequencing data (VAF information for each mutation) given underlying parameters of tumour evolution. By generating a synthetic dataset of data (x_i), parameter (θ_i), and model ($\mathcal{M}_i$)) tuples we can train hybrid variational autoencoders to jointly decompose subclonal mixtures and infer associated evolutionary parameters. Once trained, we can perform inference in real 'bulk' sequenced patient tumour biopsies.

the decomposed neutral, clonal, and potentially subclonal component distributions $k_i \in \mathbb{R}^{k \times b}$, a vector of evolutionary parameters θ_i, and the corresponding model $\mathcal{M}_i$. Consistent with previous studies, we consider models $\mathcal{M}_i$ with $i \in \{0, 1, 2\}$ where the value indicates the number of detectable subclones between 20 - 90% tumour cellular fraction (or 10 - 45% VAF when normalizing by diploid copy number).

2.3. Prior — $\theta_i, \mathcal{M}_i \sim p(\theta, \mathcal{M})$

We selected prior simulation parameter ranges consistent with previous experimental and computational estimates in tumour evolution analyses (Loeb et al., 2008; Williams et al., 2018; Werner et al., 2020; Ouellette & Awadalla, 2022). All variable parameters were uniformly sampled. A complete table of prior parameter ranges for simulating synthetic tumour VAF distributions is provided in Appendix A.2.

2.4. Hybrid variational autoencoder (HVAE)

The architecture of an HVAE includes an encoder, decoder, and an additional prediction neural network used for target label or parameter estimates. Although multiple different factorizations of the VAE can be arranged when adding a prediction network, e.g. M2-VAE Kingma et al. (2014), we opted for a simple 'linear' design for preliminary analyses (Figure 2).

Concretely, given a dataset $\mathcal{D} = \{(x_i, k_i, \theta_i, \mathcal{M}_i)\}_{1:D}$, the HVAE jointly trains an encoder with parameters ψ, a decoder with parameters ϕ, and a prediction network with parameters π to learn a latent mapping from an (i) input VAF distribution $x_i \in \mathbb{R}^{1 \times b} \rightarrow$ (ii) latent variable $z_i \rightarrow$ (iii) k component VAF distributions representing neutral 'tail', subclonal peaks, and/or clonal peak $\rightarrow$ (iv) evolutionary parameters θ_i.

Considering a single datapoint, the HVAE is trained to minimize the following objective:

$$\mathcal{L}_i(\phi, \psi, \pi; x_i) = \underbrace{-\mathbb{E}_{q_\psi(z|x_i)}[\log p_\phi(k_i|z)]}_{\text{reconstruction error}} + \underbrace{\beta \cdot \mathbb{KL}(q_\psi(z|x_i) \,||\, p(z))}_{\text{KL-divergence}} - \underbrace{\alpha \cdot \mathbb{E}_{q_\psi(z|x_i)}[\log p_\pi(\theta_i|k_i)]}_{\text{prediction error}}$$

Some important notes:

- The reconstruction error is on k_i and not the input VAF distribution x_i. This is because the goal of decoding is to learn decompose the component distributions assigned to k_i via the decoder. However, if an l1 loss is used, then the reconstruction error on x_i and k_i during training will be equivalent.
- We also include both β and α coefficients to account for large discrepancies in numerical values across VAF distribution features x_i and k_i (0 - 60000) versus target variables (normalized to 0 - 1) θ_i. This ensures that that gradients aren't biased by any given term in the objective.
- Both the prior $p(z)$ and variational distribution $q_\psi(z|x)$ were chosen to be Gaussian in this study.

Additional description of the training procedure and implementation is provided in Appendix A.2.

2.5. Datasets

To evaluate existing subclonal clustering methods, we curated or generated a total of 3 different datasets. We note that, in this study, we focus on analyzing sequencing data consisting only of VAF information from diploid copy number regions as this facilitates easier evaluation against existing methods.

2.5.1. Dataset I

Dataset I was collected from Caravagna et al. (2020). The dataset is composed of synthetic tumour sequencing data from 150 samples grown to a final population size $> 10^8$ at a birth rate of 1 and death rate of 0.2, and then sequenced to a mean depth of 120x. In terms of detectable subclones, 40 samples have no detectable subclone whereas 110 have one detectable subclone (10 - 45% VAF).

2.5.2. Dataset II

Dataset II consists of synthetic tumour sequencing data from 500 samples generated using the simulator described in 2.2 and Appendix A.1.1. Half of the samples were simulated with 0 subclones or and half were simulated with 1 subclone. Each model had 10 replicates per sequencing depth (50 - 150x) and minimum alternate reads to call a mutation (4 - 12) combination.

2.5.3. Dataset III

Dataset III was generated by directly sampling VAF information from mixture component distributions. For each replicate, $m/3$ mutations were randomly sampled from each of a Pareto distribution (shape = 20) representing the low frequency neutral 'tail' and 2 beta distributions (α, β = 50, 50) representing the subclonal (centered at 25% VAF) and clonal peaks (centered at 50% VAF). A total of 5 replicates across m = 32, 64, 128, 256, 512, 1024, 2048, 4096, 8192, 16384 were generated. This dataset was used to generate VAF information with an exactly specified number of mutations to evaluate how subclonal clustering time scales with increasing mutational burden.

2.6. Evaluated tasks

- Dataset I was used to evaluate ability for tumour evolution models to differentiate between samples with 0 or 1 detectable subclones (Task I.1).
- Dataset II was used to (Task II.1) evaluate models ability to differentiate between 0, 1, or 2 subclones at variable sequencing depths and minimum alternate read combinations, (Task II.2) accurately quantify the true subclone population fraction in the synthetic tumour, and (Task II.3) estimate the mutation rate at increasing sequencing depths.
- Dataset III was used to evaluate inference time per sample with increasing mutations (Task III.1).

3. Experiments

In addition to our proposed HVAE, we also evaluated three additional subclonal clustering methods on each of the four tasks.

- SciClone (Miller et al., 2014) is a variational Bayesian mixture model with either beta, binomial, or gaussian component distributions. We use the beta distribution variant.
- MOBSTER (Caravagna et al., 2020) is Dirichlet mixture model combining Pareto/power-law and beta distribution components. MOBSTER uses the power-law component to model the low frequency neutral tail that arises due to expanding tumour populations - this helps to avoid overestimating the number of subclones.
- TumE (Ouellette & Awadalla, 2022) is a synthetic supervised learning model that also uses simulations and deep learning to perform subclonal clustering - however, parameter inference and clustering are performed independently and only discriminative neural networks are used.

3.1. Predicting the number of subclones

In Task I.1 using synthetic tumour sequencing data generated from previous studies, we find that the proposed HVAE leads to substantial improvements in accurately quantifying the presence or absence of subclonal populations using VAF information (Table 1).

Furthermore, in Task II.2, we find that the HVAE maintains superior performance (defined by accuracy and Matthew's correlation coefficient) across all evaluated mean sequencing depth and minimum alternate reads to call a mutation combinations (Figure 3)

Table 1. Performance metrics on TASK I.1 for classifying 0 or 1 detectable subclones (MCC = Matthew's correlation coefficient).

Method	MCC	Accuracy
SciClone	0.067	37.8%
MOBSTER	0.226	62.5%
TumE	0.458	73.4%
HVAE (Ours)	**0.565**	**81.3%**

3.2. Quantifying the subclone cellular fraction

We next evaluated the ability for each method to quantify the cellular fraction of a single subclonal population, present between 10 - 45% VAF (20 - 90% cellular fraction). In this

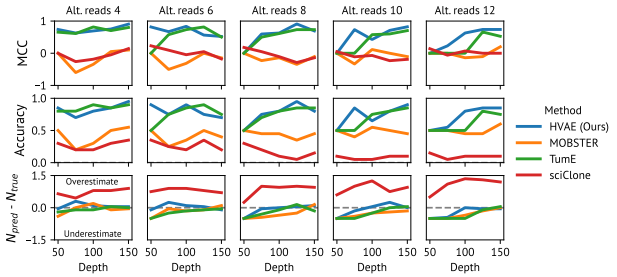

Figure 3. Performance metrics on TASK II.1 for classifying 0 or 1 subclones at variable sequencing depths (50 - 150x) and minimum alternate reads to call a mutation (4 - 12)

task, the HVAE outperformed both mixture models (SciClone and MOBSTER) but had lower performance than TumE (Table 2). However, compared to TumE which also uses a neural simulation-based inference approach, the proposed HVAE was only trained on 12.5% of the training set size as TumE (12.5 vs 40 million samples). Additional scaling of the dataset size and additional evaluation of alternative architectures will likely resolve any differences in performance on this task.

Table 2. Performance metrics on TASK II.2 for quantifying the true subclone fraction in the tumour (MAPE = mean absolute percentage error)

METHOD	MAPE
SCICLONE	34.5%
MOBSTER	19.5%
TUME	**5.6%**
HVAE (OURS)	12.5%

3.3. Predicting mutation rate

The majority of available methods do not provide evolutionary parameter estimates. However, MOBSTER provides an option to connect observed subclone frequencies and neutral 'tail' fits to analytical theory of subclonal dynamics. Although this work is preliminary, we performed a brief comparison of mutation rate estimates with MOBSTER and the HVAE (Figure 4). Estimates from both MOBSTER and HVAE improve substantially with increasing sequencing depth. This is expected as excessive sequencing noise confounds accurate assignment of mutations into each tumour population component (e.g. neutral tail, subclone, or clonal, fixed mutations), which confounds accurate mutation rate estimates.

3.4. Inference time across methods

Existing methods used for subclonal clustering tend to scale poorly in terms of compute and time with increasing mutational burden. As such, we evaluated the inference time for

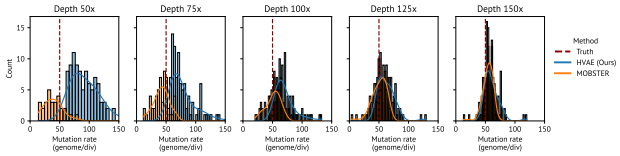

Figure 4. TASK II.3 Mutation rate estimates in 500 synthetic tumours

each method with increasing mutational burden. Compared to existing approaches, even non-mixture model approaches such as TumE, we observe substantial reductions in walltime with the HVAE (Figure 5).

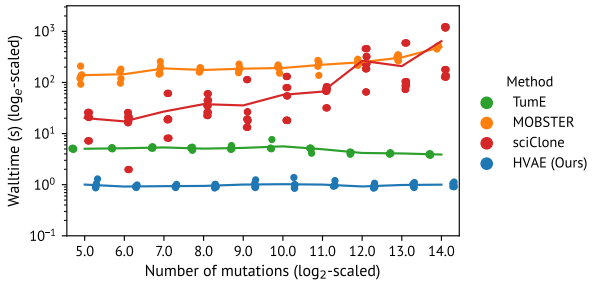

Figure 5. TASK III.1 Walltime (s) for each method with increasing 2^n mutational burden

3.5. Fitting the HVAE to real patient tumour sequencing data

In Figure 6, we provide an example fit to VAF distribution recovered from a real patient tumour biopsy.

4. Conclusion

We provide *preliminary* experiments and analyses showing that combining stochastic simulations of tumour evolution with generative neural networks, such as hybrid variational autoencoders, enables joint subclonal clustering and parameter inference in tumour populations. Evidently, it appears that jointly learning to decompose subclones and estimate evolutionary parameters improves accuracy over existing methods on the majority of tasks considered here.

However, we note that we largely focused on benchmarking subclonal clustering performance with minimal evaluation of evolutionary parameter estimates. This is largely due to a limited number of methods built that enable this type of inference. In future method comparisons, parameter estimates should clearly be better integrated into analyses, taking into account classical simulation-based inference methods such as ABC and likely more recent neural simulation-based inference methods that utilize conditional density estimators such as normalizing flows (Papamakarios & Murray, 2016;

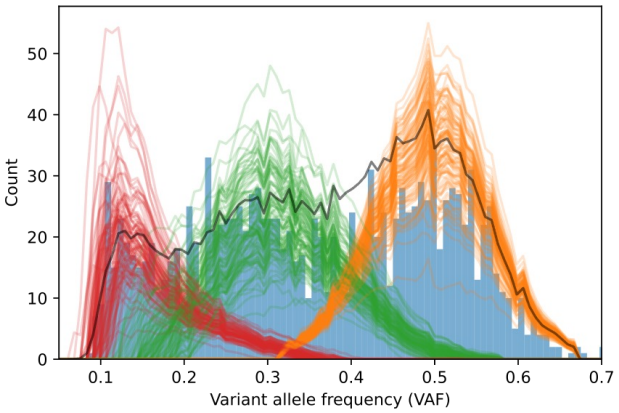

Figure 6. An example approximate fit of a VAF distribution using a trained HVAE in sequencing data from a real patient tumour biopsy. A total of 100 stochastic samples were drawn (colored) representing the neutral tail (lowest VAF; red), subclone (middle VAF; green) and clonal peak (highest VAF; orange). In addition to clustering, each fit comes with evolutionary parameter estimates (not shown here). Black represents the mean trace across all combined components.

Greenberg et al., 2019; Durkan et al., 2020) (although constraints on input/output dimensionality with invertible neural networks may pose challenges when performing clustering inference as proposed here).

In addition, although we provide accurate estimates across a variety of tasks, it isn't entirely obvious if the approximate posterior derived from repeatedly sampling the HVAE accurately captures uncertainty in a potentially multimodal ground truth posterior. Additional longer running experiments focused on sampling reference posteriors using sequential methods with theoretical guarantees (e.g. ABC) would be beneficial.

Lastly, the choice of architecture and structure of the objective function was kept simple for straightforward implementation and optimization. In this regard, evaluating a variety of alternative VAE factorizations (Kingma et al., 2014) while providing a more formal description of the objective function under joint inference of data and targets/labels would likely be justified and, ultimately, aid in interpretability and model performance.

References

Berger, M. F. and Mardis, E. R. The emerging clinical relevance of genomics in cancer medicine. *Nature Reviews Clinical Oncology*, 15, 2018. doi: 10.1038/s41571-018-0002-6.

Black, J. R. M. and McGranahan, N. Genetic and non-genetic clonal diversity in cancer evolution. *Nature Reviews Cancer*, 21, 2021. doi: 10.1038/s41568-021-00336-2.

Bozic, I. and Wu, C. Delineating the evolutionary dynamics of cancer from theory to reality. *Nature Cancer*, 1, 2020. doi: 10.1038/s43018-020-0079-6.

Bozic, I., Gerold, J. M., and Nowak, M. A. Quantifying clonal and subclonal passenger mutations in cancer evolution. *PLOS Computational Biology*, 12, 2016. doi: 10.1371/journal.pcbi.1004731.

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., and Sottoriva, A. Subclonal reconstruction of tumors by using machine learning and population genetics. *Nature Genetics*, 52, 2020. doi: 10.1038/s41588-020-0675-5.

Dentro, S. C., Wedge, D. C., and Van Loo, P. Principles of reconstructing the subclonal architecture of cancers. *Cold Spring Harbor Perspectives in Medicine*, 7, 2017. doi: 10.1101/cshperspect.a026625.

Durkan, C., Murray, I., and Papamakarios, G. On contrastive learning for likelihood-free inference, 02 2020.

Fittall, M. W. and Van Loo, P. Translating insights into tumor evolution to clinical practice: promises and challenges. *Genome Medicine*, 11, 2019. doi: 10.1186/s13073-019-0632-z.

Gillis, S. and Roth, A. Pyclone-vi: scalable inference of clonal population structures using whole genome data. *BMC Bioinformatics*, 21, 2020. doi: 10.1186/s12859-020-03919-2.

Greenberg, D., Nonnenmacher, M., and Macke, J. Automatic posterior transformation for likelihood-free inference. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2404–2414, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/greenberg19a.html.

Kingma, D. P., Rezende, D. J., Mohamed, S., and Welling, M. Semi-supervised learning with deep generative models, 2014. URL https://arxiv.org/abs/1406.5298.

Loeb, L. A., Bielas, J. H., and Beckman, R. A. Cancers exhibit a mutator phenotype: Clinical implications. *Cancer Research*, 68, 2008. doi: 10.1158/0008-5472.CAN-07-5835.

Miller, C. A., White, B. S., Dees, N. D., Griffith, M., Welch, J. S., Griffith, O. L., Vij, R., Tomasson, M. H., Graubert, T. A., Walter, M. J., Ellis, M. J., Schierding, W., DiPersio, J. F., Ley, T. J., Mardis, E. R., Wilson, R. K., and Ding, L. Sciclone: Inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLOS Computational Biology*, 10, 2014. doi: 10.1371/journal.pcbi.1003665.

Nik-Zainal, S., Loo, P. V., Wedge, D. C., Alexandrov, L. B., Greenman, C. D., Lau, K. W., Raine, K. M., Jones, D., Marshall, J. L., Ramakrishna, M., Shlien, A., Cooke, S. L., Hinton, J., Menzies, A., Stebbings, L. A., Leroy, C., Jia, M., Rance, R., Mudie, L. J., Gamble, S. J., Stephens, P. J., McLaren, S., Tarpey, P., Papaemmanuil, E., Davies, H., Varela, I., McBride, D. J., Bignell, G., Leung, K., Butler, A. P., Teague, J. W., Martin, S., Jönsson, G. B., Mariani, O., Boyault, S., Miron, P. L., Fatima, A., Langerød, A., Aparicio, S., Tutt, A. N., Sieuwerts, A. M., Borg, Å., Thomas, G. D., Salomon, A., Richardson, A. L., Børresen-Dale, A.-L., Futreal, P. A., Stratton, M. R., and Campbell, P. J. The life history of 21 breast cancers. *Cell*, 149:994 – 1007, 2012.

Ouellette, T. W. and Awadalla, P. Inferring ongoing cancer evolution from single tumour biopsies using synthetic supervised learning. *PLOS Computational Biology*, 18, 2022. doi: 10.1371/journal.pcbi.1010007.

Papamakarios, G. and Murray, I. Fast ϵ -free inference of simulation models with bayesian conditional density estimation. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 29, 2016.

Waclaw, B., Bozic, I., Pittman, M. E., Hruban, R. H., Vogelstein, B., and Nowak, M. A. A spatial model predicts that dispersal and cell turnover limit intratumour heterogeneity. *Nature*, 525, 2015. doi: 10.1371/journal.pcbi.1004731.

Werner, B., Case, J., Williams, M., Chkhaidze, K., Temko, D., Fernández-mateos, J., Cresswell, G., Nichol, D., Cross, W., Spiteri, I., Huang, W., Tomlinson, I., Barnes, C., Graham, T., and Sottoriva, A. Measuring single cell divisions in human tissues from multi-region sequencing data. *Nature Communications*, 11, 2020. doi: 10.1038/s41467-020-14844-6.

Williams, M. J., Werner, B., Heide, T., Curtis, C., Barnes, C. P., Sottoriva, A., and Graham, T. A. Quantification of subclonal selection in cancer from bulk sequencing data. *Nature Genetics*, 50, 2018. doi: 10.1038/s41588-018-0128-6.

A. Appendix

A.1. Additional information on the synthetic data generation process

A.1.1. Simulation algorithm

The stochastic branch process of tumour growth and evolution follows a variant of the Gillespie algorithm (Williams et al., 2018). Here, a tumour grows with each cell dividing or dying with probabilities proportional to the birth rate b or death rate d, respectively. During each division, a cell will acquire a Poisson-distributed number of mutations based on the mutation rate μ. If a cell acquires a driver mutation, its overall growth rate $(b - d)$ is scaled by a factor of $(1 + s)$ where s indicates the selection coefficient for the driver mutation. Once a tumour reaches a specified final population size N, the tumour is virtually sequenced under a binomial or beta-binomial sequencing noise model that is consistent with the empirical data generation process. This can be more formally denoted within a pseudo-algorithm.

Algorithm 1 Pseudo-algorithm for simulating tumour evolution and synthetic tumour sequencing data

Set the detectable subclone frequency range (f_{min}, f_{max}) for simulations with ≥ 0 subclones
Sample parameters from prior $\theta_i, \mathcal{M}_i \sim p(\theta, \mathcal{M})$
while ($f_{min} >$ subclone frequency $> f_{max}$) & (number of subclones $\neq \mathcal{M}_i$) **do**
  Set current population size $n = 1$
  Set time $t = 0.0$
  Initialize founder cell with n_{clonal} clonal mutations
  while n $< N$ **do**
    Randomly sample a cell j from population
    Sample $r \sim$ Unif(a,b) where a = 0 and b = b_{max} + d_{max}
    if $b_j > r$ **then**
      Cell divides and both daughter cells acquire k mutations where k is Poisson distributed with mean equal to the per genome division mutation rate μ
      $n = n + 1$
    else if b_j + $d_j > r >= b_j$ **then**
      Cell dies
      $n = n - 1$
    else
      Nothing happens
    end if
    $\tau \sim$ Exp(1) or -log(Unif(0, 1))
    $t = t + \tau n^{-1}(b_{max} + d_{max})^{-1}$
  end while
end while
Virtual biopsy synthetic tumour

The virtual biopsy procedure follows the expected variation seen in next-generation sequencing data. Namely, given a tumour with population size N_t and mean sequencing depth $\bar{d}$, the total observed read depth d_i covering each mutation i is distributed as $d_i \sim BetaBin(n = N_t, p = \bar{d}/N_t, \rho)$. Here, ρ indicates the overdispersion parameter for the Beta distribution. When ρ is zero, d is is binomially distributed. Then, given the total read depth d_i, the true variant allele frequency VAF_{t_i}, and the tumour purity ϕ, the alternate read count r_i for mutation i is distributed as $r_i \sim Binomial(n = d_i, p = VAF_{t_i})$. Given r_i, the observed noisy VAF_{o_i} for each mutation i is computed as r_i/d_i.

A.1.2. Simulation parameters

For simulation parameter selection, we follow similar specifications as previous work (Williams et al., 2018; Ouellette & Awadalla, 2022). Most notably, we take advantage of the fact that VAF distributions do not encode information on tumour population size (Williams et al., 2018). This means we can simulate smaller population sizes to take advantage of the reduced computational burden and improved simulation speed. In addition, we fix the death rate to 0 as it also improves computational efficiency. We do note that previous work has shown that the ratio of birth and death rates may lead to potentially detectable deviations in the shape of the VAF distribution (Bozic et al., 2016). However, we do not consider

that problem here, as it has been previously shown that simulation-based cancer evolution inference methods are robust to changes in the underlying birth and death rate combinations (Ouellette & Awadalla, 2022). Below we list the variable and fixed parameters used to generate synthetic tumour sequencing data.

Table 3. Prior parameter ranges for tumour evolution simulations

PARAMETER	MIN	MAX	SAMPLING
MUTATION RATE (GENOME/DIVISION)	1	500	UNIFORM
NUMBER CLONAL MUTATIONS	1	5000	UNIFORM
SUBCLONE FITNESS (1+S)	1	24	UNIFORM
SUBCLONE EMERGENCE TIME (% N_{final})	0.002	0.5	UNIFORM
SEQUENCING DEPTH	50	200	UNIFORM
SEQUENCING OVERDISPERSION (ρ)	0.0	0.01	UNIFORM
MINIMUM ALTERNATE READS TO CALL MUTATION	4	12	UNIFORM
BIRTH RATE	$log(2)$	$log(2)$	FIXED
DEATH RATE	0.0	0.0	FIXED
FINAL TUMOUR SIZE (N_{final})	1000	1000	FIXED

A.2. Details on HVAE training and implementation

We simulated 5 million samples under each model (0, 1, or 2 subclones) described in the simulator above. Input VAF distributions x_i, and each component distribution within k_i, were histograms with b = 100 bins generated by tabulating all mutations from 0 to 70% VAF. Using this synthetic dataset, we then trained four different HVAE models in total:

- HVAE_{ms} was used to perform model selection across each evolutionary model (0, 1, or 2 subclones)
- HVAE_0 was used to perform VAF distribution clustering and parameter estimates under the 0 subclone evolutionary model ($\mathcal{M}_0$)
- HVAE_1 was used to perform VAF distribution clustering and parameter estimates under the 1 subclone evolutionary model ($\mathcal{M}_1$)
- HVAE_2 was used to perform VAF distribution clustering and parameter estimates under the 2 subclone evolutionary model ($\mathcal{M}_2$)

For each HVAE, the encoder, decoder, and prediction networks were fully-connected multi-layer perceptrons implemented in pytorch. To select optimal configurations for each neural network, we performed random hyperparameter search 150 times with the following hyperparameter ranges:

Table 4. Hyperparameter ranges used in random search

HYPERPARAMETER	MIN	MAX
HIDDEN NEURONS	256	1024
HIDDEN LAYERS	3	7
LATENT DIMENSIONS	2	12
ACTIVATION FUNCTIONS	GELU, RELU,HARDSWISH,LEAKY RELU	
RECONSTRUCTION LOSS	L2, L1, SMOOTH L1	
PREDICTION LOSS	L2, L1, SMOOTH L1	
ALPHA	1	100
BETA	0.5	1.5
DROPOUT	0	0.5
LEARNING RATE	10^{-10}	10^{-3}
WEIGHT DECAY (ADAM OPTIMIZER)	0	0.2
BATCH SIZE	256	1024