LiquidBayes: A Bayesian Network for Monitoring Cancer Progression Using Liquid Biopsies

by

Kevin Yang

BSc. Computer Science and Statistics, The University of British Columbia, 2020

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Bioinformatics

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Bioinformatics)

The University of British Columbia
(Vancouver)

December 2022

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the thesis entitled:

LiquidBayes: A Bayesian Network for Monitoring Cancer Progression Using Liquid Biopsies

submitted by **Kevin Yang** in partial fulfillment of the requirements for the degree of **Master of Bioinformatics** in **Bioinformatics**.

Examining Committee:

Andrew Roth, Assistant Professor, Computer Science, UBC
Supervisor

Yongjin Park, Assistant Professor, Statistics, UBC
Co-supervisor

Alexandre Bouchard-Côté, Professor, Statistics, UBC
Supervisory Committee Member

Ryan Morin, Associate Professor, Molecular Biology & Biochemistry, SFU
Supervisory Committee Memeber

Abstract

Cancer exhibits temporal heterogeneity, described by the existence and continual evolution of multiple cell subpopulations in a tumour. Temporal heterogeneity triggers resistance, disrupting targeted therapies and worsening patient prognosis. The continual monitoring of cancer patients can aid in identifying resistance, leading to informed treatment decisions. Liquid biopsies are a non-invasive blood sample containing double stranded, unbound DNA called cell-free DNA (cfDNA). A subset of cfDNA, known as circulating tumour DNA (ctDNA), originates from the tumour itself and can uncover tumour-specific mutations to characterize cancer. Although liquid biopsies provide a means to evaluate therapy response at multiple time points, it is known that ctDNA abundance is extremely low, especially in post-treatment patients.

Statistical methods have been developed for estimating tumour burden using ctDNA samples. Moreover, some groups have integrated somatic mutations derived from bulk sequencing of a tissue biopsy to address low ctDNA abundance. However, no method we know of incorporates single-cell sequencing of a tissue biopsy in ctDNA analysis. In this thesis, we present LiquidBayes, a Bayesian Network that integrates clone-level copy number profiles from single-cell Whole Genome Sequencing (WGS) of the primary tissue with WGS of ctDNA samples. LiquidBayes leverages Markov Chain Monte Carlo techniques and employs Probabilistic Programming Languages, gaining access to efficient sampling methods. LiquidBayes significantly outperforms state-of-the-art methods in tumour burden estimation and additionally allows for characterizations of clonal prevalences. LiquidBayes offers the ability to analyze serial ctDNA samples to dissect temporal heterogeneity, intercept resistance and improve patient prognosis.

Lay Summary

Cancer is comprised of subpopulations of cells, known as clones, each of which may respond differently to treatments. As such, it is important to monitor tumour progression after administering therapy to inform subsequent treatment decisions. Liquid biopsies are non-invasive blood draws that contain circulating-tumour DNA (ctDNA), DNA fragments derived from the tumour itself. By studying ctDNA, we can uncover salient properties of cancer; in particular, tumour burden (pervasiveness of cancer) and clonal prevalences (diversity of cancer).

Various methods have been developed to estimate tumour burden using ctDNA samples. However, these methods do not analyze cancer at single-cell resolution. Our objective is to incorporate a tissue biopsy, analyzed at single-cell resolution, with liquid biopsies. We present LiquidBayes, a Bayesian statistical model for inferring tumour burden and clonal prevalences, leading to informed treatment decisions and improved prognosis.

Preface

This thesis was completed under the supervision of Dr. Andrew Roth and co-supervision of Dr. Yongjin Park at the BC Cancer Research Centre. I was responsible for implementing LiquidBayes and all preprocessing steps for experiments. Felix Fu joined as a summer research student and contributed to the development of the benchmarking pipeline. Shaocheng Wu was the provider of the processed Lymphoma Dataset described in Section 2.3.1.

Table of Contents

List of Tables

List of Figures

Glossary

DNA Deoxyribonucleic Acid

CFDNA cell-free DNA

CTDNA circulating tumour DNA

NGS Next-Generation Sequencing

WGS Whole Genome Sequencing

MRD Minimal Residual Disease

BN Bayesian Network

CPD Conditional Probability Distribution

MCMC Markov Chain Monte Carlo

TFRI Terry Fox Research Institute

FL Follicular Lymphoma

DLBCL Diffuse Large B-Cell Lymphoma

CN copy number

CNV Copy Number Variation

VCF Variant Call Format

SNV Single Nucleotide Variant

GMM Gaussian Mixture Model

PPL Probabilistic Programming Language

SOTA state-of-the-art

VAF Variant Allele Frequency

HDI Highest Density Interval

HDR Highest Density Region

HMM Hidden Markov Model

DLP+ Direct Library Preparation Plus

DAG Directed Acyclic Graph

PCR Polymerase Chain Reaction

SNP Single Nucleotide Polymorphism

Acknowledgments

I would like to thank my supervisor Dr. Andrew Roth for his guidance throughout the past two years. He was available to answer questions and give feedback to drive this project forward. Many thanks to my co-supervisor, Dr. Yongjin Park for supporting me in academic and family matters. I am grateful to Dr. Andrew Roth and Dr. Yongjin Park for accommodating me as I welcomed my daughter Felicity into this world.

I also want to acknowledge those part of my committee. Thank you Dr. Ryan Morin and Dr. Alexandre Bouchard-Côté for providing valuable input during committee meetings. This thesis has improved as a result. I also want to extend my appreciation to Dr. Paul Pavlidis for agreeing to chair my thesis defense. This thesis would not be complete without his contribution.

A big thanks to the Canadian Institute of Health Research for partially funding my research through the Canada Graduate Scholarships Canada Graduate Masters award.

I am grateful for the opportunity to work in the Roth lab for these past two years, where I have developed both as an individual and as a researcher. Meaningful friendships have made this journey more enjoyable.

Above all, thank you to my beautiful wife Cheyenne whom I cherish with all my heart. This thesis simply would not exist without your hard work and perseverance assuming familial responsibilities on top of your own research, studies and personal life. Finally, I would be remiss not to mention my wonderful daughter Felicity, who keeps everybody up all night changing diapers, feeding and cleaning bottles. You bring much joy to our family.

Chapter 1

Introduction

Cancer is a complex disease driven by genetic mutations. Originating from a single mutated cell, subsequent rounds of proliferation and additional mutations ultimately give rise to the tumour[46]. This evolutionary process produces genetically distinct populations of cells known as 'clones', leading to inconsistent therapeutic response across the tumour[45][50]. Clonal evolution exhibits temporal heterogeneity, marked by the growing or waning of clones over time due to selective pressures. Importantly, clones in cancers develop resistance to treatment by acquiring mutations that alter cell-intrinsic mechanisms which govern its response to therapies[63][50]. Consequently, continual monitoring at all stages of treatment can improve prognosis by intercepting clonal resistance. Currently, clonal analysis is commonly done using invasive tissue biopsies. However, tissue biopsies are unfit for monitoring treatment response, as it is impractical to extract multiple tissue biopsies across time. In contrast, liquid biopsies are non-invasive blood samples containing double stranded, unbound Deoxyribonucleic Acid (DNA), known as cell-free DNA (CFDNA). In cancer patients, circulating tumour DNA (CTDNA) is a subset of CFDNA originating from the tumour. Studies have shown high concordance between CTDNA and tumour biopsies[7][1][13][23], suggesting CTDNA to act as a proxy for serial tissue biopsies for monitoring cancer progression. In short, tracking the movement of clones using CTDNA will enable clinicians to make informed treatment decisions for patients.

1.1 Tumour Heterogeneity

Cancer is uncontrolled cell growth which follows evolutionary principles, where genetic variation alters molecular signatures in cells[46]. Therefore, the tumour is composed of genotypically distinct subpopulations of cells called clones – this condition is known as tumour heterogeneity. Tumour heterogeneity has clinical relevance, in that targeted therapies do not have a uniform effect on the tumour. Incidentally, cancer can develop resistance as tumour cells accrue mutations that alter genetic pathways and therapy targets[50]. Hence, tumour heterogeneity introduces a great deal of complexity when developing effective therapies[9]. Sequencing technologies have the potential to uncover tumour heterogeneity, monitor clonal fluctuation and identify the emergence of clinical resistance[3] (Section 1.3). Furthermore, genetic mutations such as Copy Number Variations (CNVS) (Section 1.4.1) and Single Nucleotide Variants (SNVS) (Section 1.4.2) can be used to characterize tumour heterogeneity. Statistical models have successfully applied Next-Generation Sequencing (NGS) data for analyzing tumour heterogeneity[40][21][51][54][47].

1.2 Liquid Biopsies

Tissue biopsy is the removal of tissue from a patient to provide a representative specimen for interpretation and analysis[58]. Derived from the tumour itself, tissue biopsy is the gold standard source of data in cancer research[26]. However, a tissue biopsy cannot portray cancer holistically. Tissue biopsies suffer from three major drawbacks: first, sampling bias is present when extracting tissue; second, spatial heterogeneity limits its representative scope; and lastly, temporal heterogeneity is ignored as serial sampling is infeasible [55]. These limitations indicate that tissue biopsies are insufficient to comprehensively portray cancer.

Mandel and Metais first reported the existence of fragmented DNA in the non-cellular component of blood[48]. A liquid biopsy is a non-invasive sample of biological fluid containing fragmented DNA called CFDNA. CFDNA are double stranded, highly fragmented molecules that are approximately 150bp in length[62]. In cancer patients, circulating tumour DNA (CTDNA) is a subset of CFDNA that originates from the primary tumour. Importantly, studies have illustrated a high

concordance between CTDNA and tumour biopsies[7]. Therefore, CTDNA reveals relevant characteristics of the tumour for therapy trials. Applications of CTDNA are numerous: diagnosis and molecular profiling, tracking of therapeutic response, monitoring resistance, studying tumour heterogeneity, detecting Minimal Residual Disease (MRD) and early cancer detection[7]. However, CTDNA is present in very low proportions, impeding its utility in clinical contexts.

Presently, there are several statistical methods for analyzing ctDNA. ichorCNA[1] and LiquidCNA[35] use copy number (CN) to track and quantify clonal evolution and prevalence, whereas Kang et al.[27] and Li et al.[37] propose methods that leverage CTDNA methylation patterns. Instead of focusing exclusively on CTDNA, Zviran et al.[69] introduce an integrated bulk analysis of ctDNA and solid tissue to address low CTDNA abundance. However, bulk approaches are inadequate to resolve minor clonal populations due to sequencing error rates[19] and fail to address clonal CNVS at low tumour cellularity[14][66]. Considering these limitations, we employ Direct Library Preparation Plus (DLP+)(Section 1.3.2), delivering single-cell resolution for precise deconvolution of both major and minor clonal populations and resolving clonal CNVS in low tumour burden settings.

1.3 Sequencing Technologies

1.3.1 Single-Cell Sequencing

NGS allows for the simultaneous sequencing of millions of different DNA molecules. NGS has substantially increased accessibility and speed of sequencing, allowing researchers to uncover genetic alterations that underlie the pathogenesis of cancer at an unprecedented scale[43]. Furthermore, DNA sequencing has progressed in precision and throughput, allowing for sequencing of entire genomes of individual cells[53]; this methodology is referred to as single-cell sequencing. Single-cell sequencing can reveal genomic variability among individual cells, delivering unprecedented insights into tumour heterogeneity. In particular, it enables identification of minor clonal populations and reconstruction of clonal phylogenies. However, single-cell sequencing methods traditionally rely on genome amplification, which leads to uneven coverage and allelic dropout[67][44]. We tackle this

issue by employing a unique library preparation method, DLP+, to mitigate these biases.

1.3.2 Direct Library Preparation+

DLP+ is a scalable single-cell library preparation method without preamplification[36]. DLP+ distinguishes itself by performing shallow sequencing of thousands of cells rather than deep sequencing of few cells. In DLP+, object recognition is used to assess cell state, quality and doublets. A tagmentation step appends unique oligonucleotide barcodes to exposed DNA in each well for mapping reads back to their respective cells. Then, rounds of Polymerase Chain Reaction (PCR) are performed on individual wells. DLP+ identifies clonal populations by clustering cells on their CN profiles. Briefly, UMAP[39] is applied to normalized raw copy number data from HMMcopy[34]. Next, the reduced data are clustered, where clusters represent clones. Then, cells in each cluster are merged to produce clone-level pseudo-bulk genomes. Figure 1.1 gives a high level overview of the DLP+ pipeline.

1.4 Cancer Genomics

1.4.1 Copy Number Variation

CNVS are gains or deletions of genomic segments and account for a substantial proportion of human genetic variations[64]. Customarily, CNVS have been defined to be "a segment of DNA that is 1kb or larger and is present at a variable copy number in comparison with a reference genome."[15]. Significantly, CNVS are prevalent in cancer and can elucidate causative biological mechanisms and impart prognostic insights for patients[57]. Additionally, CN profiles are an effective way of summarizing genome-wide CNVS. CN profiles are a set of integers representing the CN at each bin, where bins are non-overlapping segments of the genome with fixed length. Statistical methods that leverage CN information for interrogating tumour heterogeneity have displayed fruitful results[21][54][40][16][47].

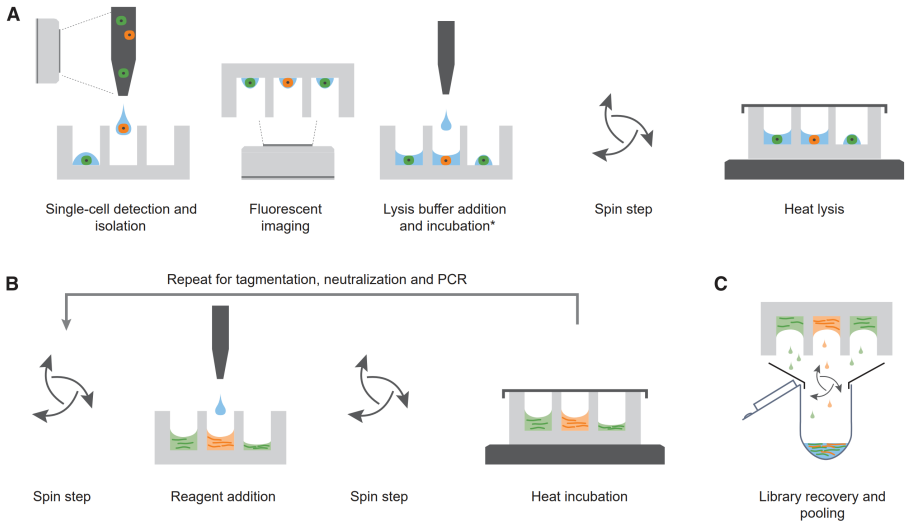


Figure 1.1: Overview of the DLP+ library preparation methodology. **A:** First, 1000s of single cells are isolated into wells on a chip, imaged, then lysed. **B:** DNA undergoes tagmentation, heat incubation and rounds of spinning. **C:** Genetic material from wells are pooled together in preparation for sequencing. (Figure taken from [36])

1.4.2 Single Nucleotide Variant

SNVS describe the event where a single nucleotide is substituted at a specific genomic position[68]. This substitution can alter amino acid synthesis, disrupt protein function, and ultimately effect disease. In cancer, variants may confer selective advantages to a subpopulation of cells, promoting tumour growth[60]. These variants are called 'driver mutations'. An active area of research examines driver mutations for the purpose of prognosis evaluation and treatment response[59]. Variant calling, the process whereby genomic positions that harbor mutations are discerned, is a common component in bioinformatics pipelines. Numerous variant callers have been developed, each with their own strengths and weaknesses[65].

1.5 Probabilistic Models

1.5.1 Bayesian Networks

Bayesian Networks (BNS) (or graphical models) are Directed Acyclic Graphs (DAGS) representing probability distributions [30]. Nodes are associated with statistical distributions (e.g. Gaussian) and edges express interactions between nodes. They visualize the structure of a probabilistic model by defining conditional independence properties[5]. Each node has a Conditional Probability Distribution (CPD), which is a function of its parents. In this way, the nodes in a BN are linked through probabilistic associations. Typically, BNS are used in conjunction with Bayesian inference, facilitating model design and inference method selection.

1.5.2 Forward Sampling

Forward sampling is a method for sampling a BN. Nodes are sampled in an order such that upon sampling a node, values for all of its parents exist[29]. In this setting, we sample each node using its CPD. The set of samples for all nodes is called a particle. Particles can then be used to estimate expectations and probabilities.

1.5.3 Markov Chain Monte Carlo

A common issue in Bayesian inference deals with the intractable normalization constant in the posterior distribution. Given a prior $p(\theta)$ and likelihood $p(x \mid \theta)$ over observations x, the posterior distribution is defined as,

$$p(\theta \mid x) = \frac{p(\theta)p(x \mid \theta)}{\int p(x \mid \theta)p(\theta)\,d\theta} \tag{1.1}$$

Monte Carlo simulation draws i.i.d samples from a non-standard target distribution $p(x)$ defined on a high-dimensional space. Empirical point estimates can be used to approximate the target distribution[2]. Specifically, if we have N samples,

$$p_N(x) = \frac{1}{N}\sum_{i=1}^{N} \delta_{x^{(i)}}(x) \tag{1.2}$$

where $\delta_{x^{(i)}}(x)$ is the delta-Dirac mass located at $x^{(i)}$, specifies an estimate of the mean of the posterior distribution. Markov Chain Monte Carlo (MCMC) is a way to sample from complicated probability distributions in high dimensions. MCMC techniques such as Metropolis Hastings[22][41] and Gibbs Sampling[18][17] are particularly useful in approximating intractable integrals in Bayesian inference.

1.5.4 Bayesian Coverage

The Bayesian coverage is the proportion of runs whose Highest Density Interval (HDI) contains the true parameter value. Hyndman [25] defines the HDI (or Highest Density Region (HDR)) in the following manner. Let $f(x)$ be the density function of a random variable X. Then the $100(1-\alpha)\%$ HDR is the subset $R(f_\alpha)$ of the sample space of X such that

$$R(f_\alpha) = \{x : f(x) \geq f_\alpha\} \tag{1.3}$$

where f_α is the largest constant such that $Pr(X \in R(f_\alpha)) \geq 1-\alpha$.

Chapter 2

Methods

2.1 LiquidBayes

2.1.1 Models

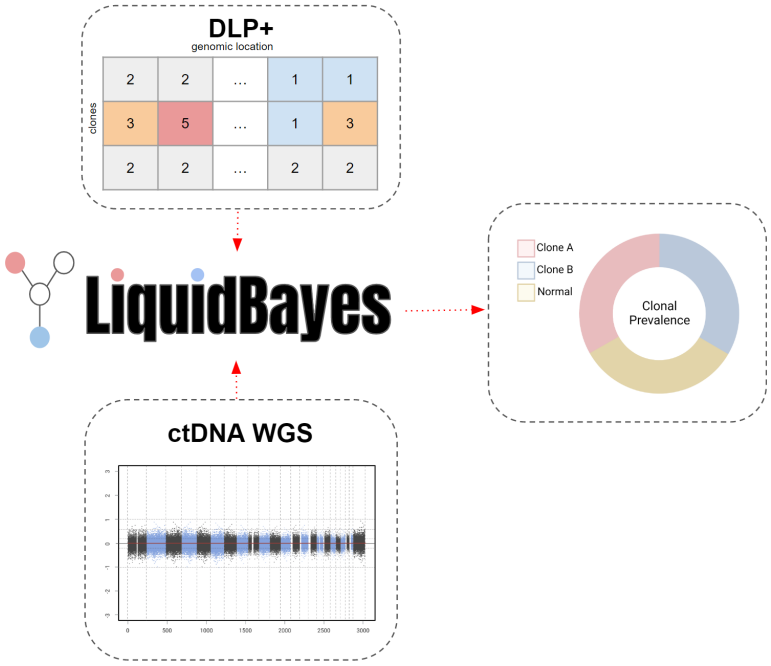
LiquidBayes was implemented as a BN and offered two models - a base model and an extended model.

Base Model

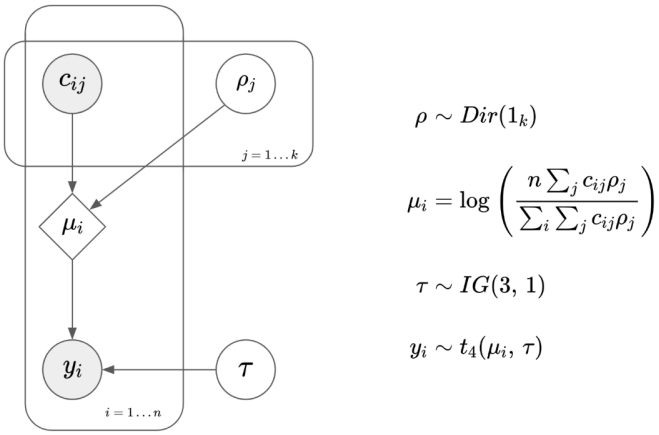
The base model leveraged clone-specific CN profiles to enhance tumour fraction and clonal prevalence estimates. A general schematic and its graphical model is shown in Figure 2.1. We modeled y_i, transformed binned read counts, using a Student's t-distribution with mean parameter μ_i and scale parameter τ. μ_i was the log-transform of the linear combination between bin-specific clonal CN values and clonal prevalences, divided by the sample ploidy estimate. τ was modeled as an Inverse-gamma(3,1) and ρ a Dirichlet using a vector of 1's with length k.

Extended Model

The extended model incorporated SNVS alongside clone-specific CN profiles. A general schematic and its graphical model is shown in Figure 2.2. The portion related to clone-specific CN profiles was the same as the base model (Section 2.1.1).



(a) General schematic for LiquidBayes' base model. Inputs include per-clone CN profiles from DLP+ and binned read counts from Whole Genome Sequencing (WGS) of ctDNA. The model outputs clone-specific and normal prevalence estimates.



(b) Graphical model for LiquidBayes base model. k is the number of clones and n is the number of bins.

Figure 2.1: LiquidBayes base model.

We estimated m_{lj}, the number of mutant copies at site l for clone j, by multiplying the Variant Allele Frequency (VAF) at that site ($\frac{f'_{lj}}{f_{lj}}$) with the corresponding CN value for the bin containing the site (c_{qj}). We modeled the number of mutant reads from CTDNA at site l (b_l) using the Binomial Logit distribution, where the number of trials was the total number of reads (d_l) and logits were $\xi_l = \sum_j \rho_j m_{lj}$, ρ being the clone proportions. SNVS were treated as biallelic.

2.1.2 Preprocessing

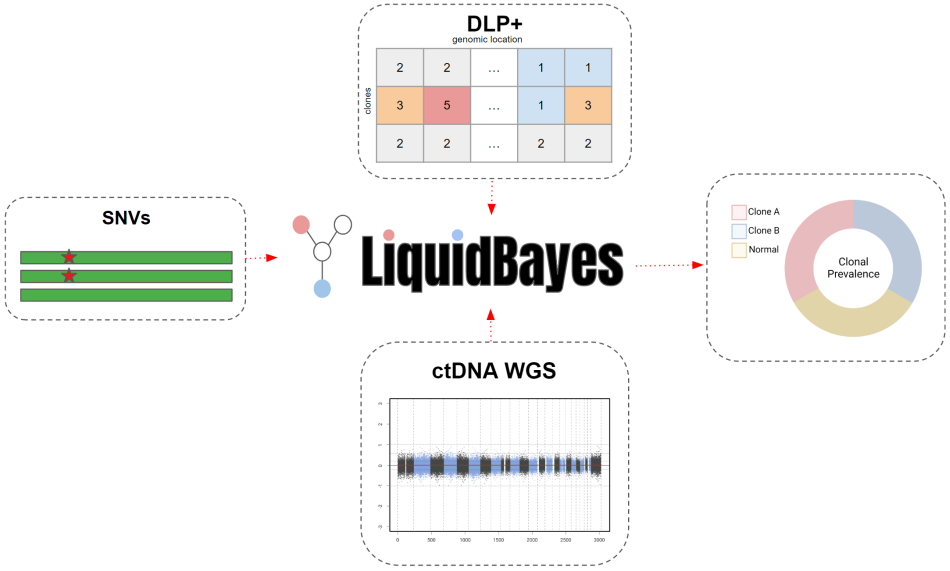
Read Counts

LiquidBayes included a preprocessing pipeline to correct and normalize raw read counts. Binned read counts were extracted from the CTDNA bam using `readCounter` from hmmcopy_utils[33]. GC and mappability bias correction were applied using `correctReadcount` from HMMcopy[34]; GC and mappability wig files were generated using `gcCounter` and `mapCounter` from hmmcopy_utils[33]. The `copy` column of the resultant dataframe contained normalized and corrected binned read counts. The hg19 reference genome was used in our experiments. Figure 2.3 displays the effect of correction and normalization at different tumour fractions.

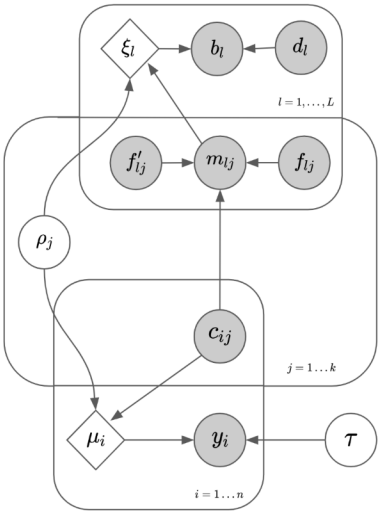
Subsequently, we removed outliers from corrected and normalized binned read counts. We noted that the number of read counts across bins with identical bin-specific CN configurations should be normally distributed. Therefore, we grouped read counts according to bin-specific CN configurations; for each group, we fitted a two component Gaussian Mixture Model (GMM) and removed read counts in the component with the higher covariance. Bin-specific CN configurations were ordered tuples of clonal and normal CN values. Possible CN configurations for a setting with two clones and a matched normal were $\{$(2,2,2), (2,2,3), (4,3,2)$\}$. Figure 2.4 illustrates the data before and after removing outliers using this method.

Copy-Number Profiles

We performed no additional preprocessing steps for CN profiles.



(a) General schematic for LiquidBayes extended model. Inputs include per-clone CN profiles from DLP+, binned read counts from WGS of ctDNA and SNVS from DLP+ and ctDNA.



$$\rho \sim \text{Dir}(1_k)$$

$$\tau \sim \text{IG}(3,\, 1)$$

$$\mu_i = \log\left(\frac{n \sum_j c_{ij}\rho_j}{\sum_i \sum_j c_{ij}\rho_j}\right)$$

$$m_{lj} = \frac{f'_{lj}}{f_{lj}} c_{qj} \text{ where } l \in \text{bin } q$$

$$\xi_l = \sum_j \rho_j m_{lj}$$

$$b_l \mid d_l,\, \xi_l \sim \text{BinLogits}(d_l,\, \xi_l)$$

$$y_i \mid \mu_i,\, \tau \sim t_4(\mu_i,\, \tau)$$

(b) Graphical model for LiquidBayes extended model. k is the number of clones, n is the number of bins and L is the number of shared SNV sites across all clones.

Figure 2.2: LiquidBayes extended model.

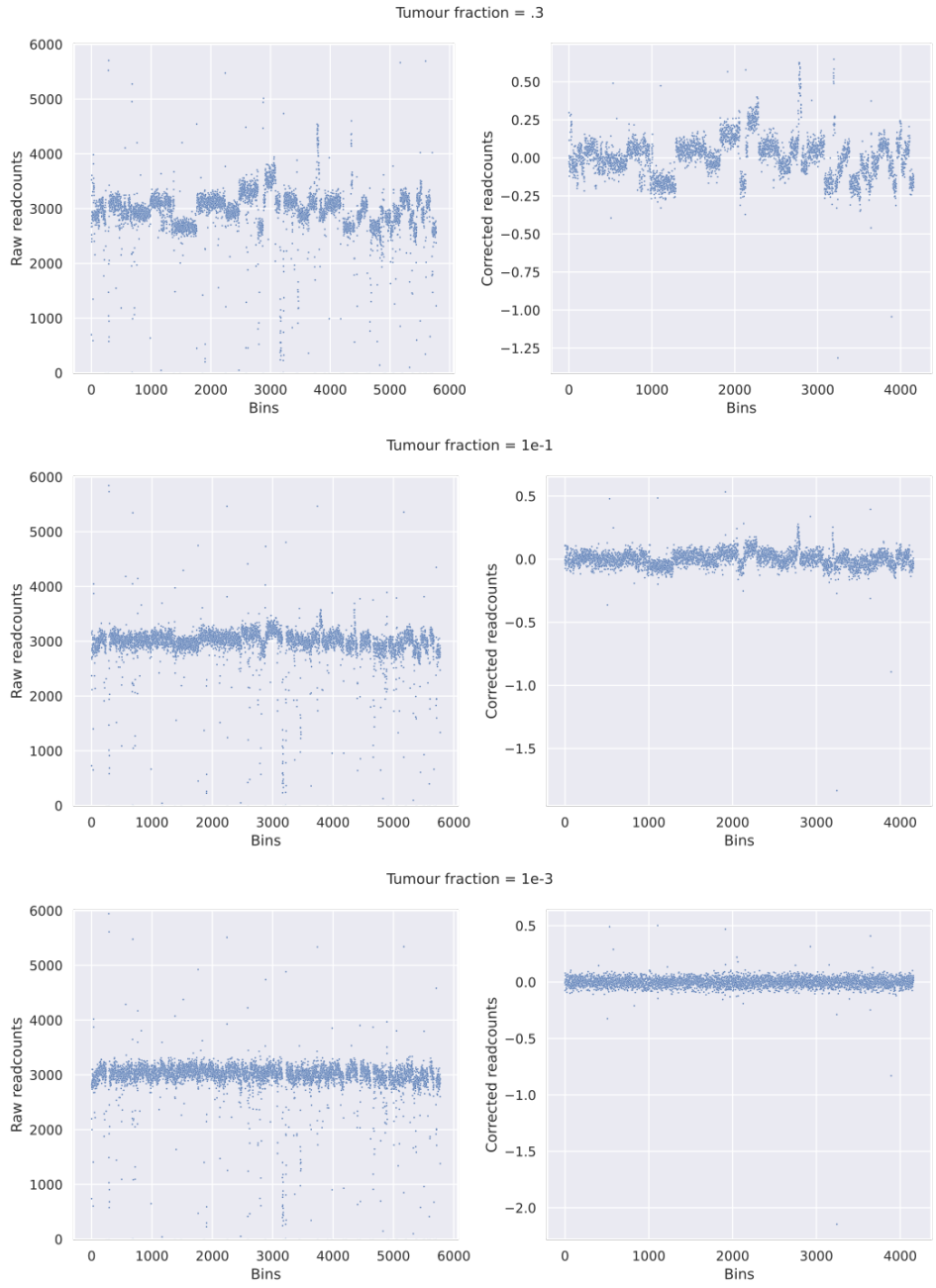


Figure 2.3: Raw and corrected read counts at different tumour fraction levels. **Left:** Raw read counts from `readCounter`. **Right:** Read counts after `correctReadcount` was applied.

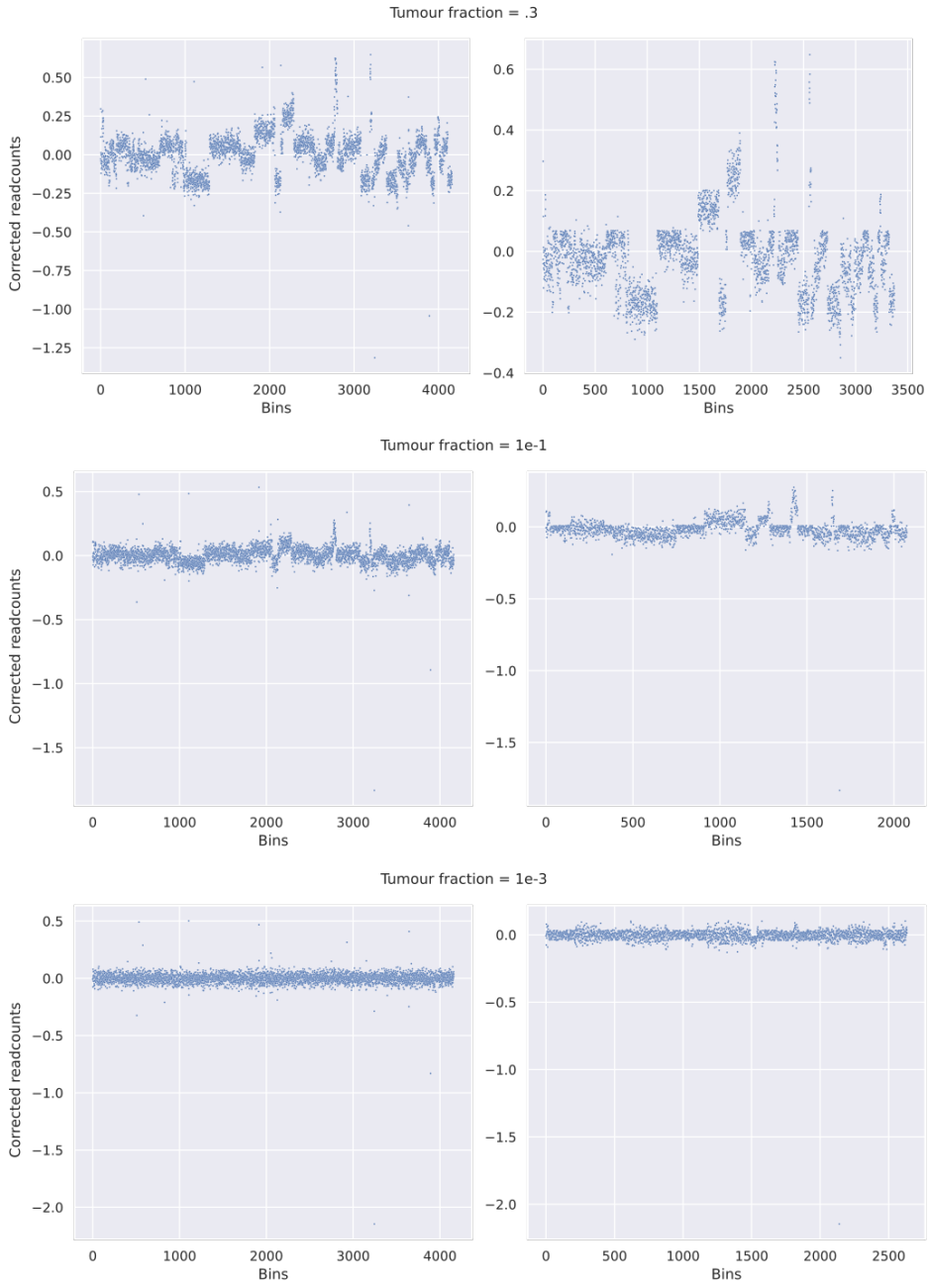


Figure 2.4: Corrected read counts before and after removing outliers at different tumour fractions. **Left:** Corrected read counts. **Right:** Read counts after outliers were removed.

Single Nucleotide Variants

SNV preprocessing only applied to the extended model. First, we obtained reference and alternate allele counts for all clones and the CTDNA sample, giving $K+1$ Variant Call Format (VCF) files, K being the number of clones. Next, we filtered duplicate sites, sites with no counts and sites not present in any CN profile bin. We computed clone-specific VAFS at the union of sites across all clones, imputing missing values (VAF of a clone not harboring an SNV at the site) with zero. Then, we multiplied clone-specific VAFS by the corresponding clone-specific CN value to produce clone-specific mutant copy estimates. Finally, we constructed an $L \times (2+K)$ `ndarray`, where L was the number of sites post-filtering and K was the number of clones. The first two columns were the reference and alternate allele counts from the CTDNA sample and the remaining K columns were the clonal mutant copy estimates.

2.1.3 Implementation

LiquidBayes was implemented using the `numpyro` Probabilistic Programming Language (PPL)[49]. We applied the `numpyro.infer.MCMC` kernel and used the `numpyro.infer.NUTS` sampler[24]. For `numpyro.infer.MCMC`, we set `num_samples=10000` and `num_warmup=500`. For `numpyro.infer.NUTS`, we set `target_accept_prob=.95`. All other parameters for both functions remained as their default values. Two `numpyro` model functions were constructed corresponding to each version of LiquidBayes. LiquidBayes was wrapped as a command line interface program using the `click` Python package. Source code can be found at (https://github.com/Roth-Lab/LiquidBayes)

2.2 Synthetic Datasets

2.2.1 Simulating Copy Number Profiles

Synthetic CN profiles were simulated using an Hidden Markov Model (HMM). We imposed 8 distinct states which corresponded to CN states 0-8. Each state had a self-loop and positive transition probabilities for two steps in either direction (e.g. state 3 can transition to one of 5 states: 1,2,3,4,5). We initialized the first CN state

to 2. Given the current state, we sampled from a Multinomial with transition probabilities $[.005, .01, .97, .01, .005]$ using `numpy.multinomial`. Next, we indexed the vector $[-2, -1, 0, 1, 2]$ using the sampled value and added the indexed value to the current state to get the next state. The emission distribution was the identity. The upper and lower bounds for the CN state were 8 and 0, respectively. Each synthetic CN profile contained 5000 bins and we treated CN profiles as independent. Algorithm 1 presents pseudocode for this process. Figure 2.5 depicts simulated CN profiles for 3 clones.

Algorithm 1: Simulate Copy-Number Profiles

```
p ← [.005, .01, .97, .01, .005]    // transition probabilities
events ← [−2, −1, 0, 1, 2]                          // events
cn_profiles ← []
for i in range(k) do
    cn_profile ← [2]
    for j in range(1, n) do                  // n =number of bins
        idx ← Multi(1, p)
        state ← events[idx] + cn_profile[j − 1]

        // force values to be in [0,8]
        if state > 8 then
            state ← 8
        end
        else if event < 0 then
            state ← 0
        end
        cn_profile.append(state)
    end
    cn_profiles.append(cn_profile)
end
```

For experiments involving missing clones, we either removed the clone with the largest or smallest proportion. We accomplished this by deleting the appropriate CN from the original set of CN profiles and writing the updated set of CN profiles to a new file.

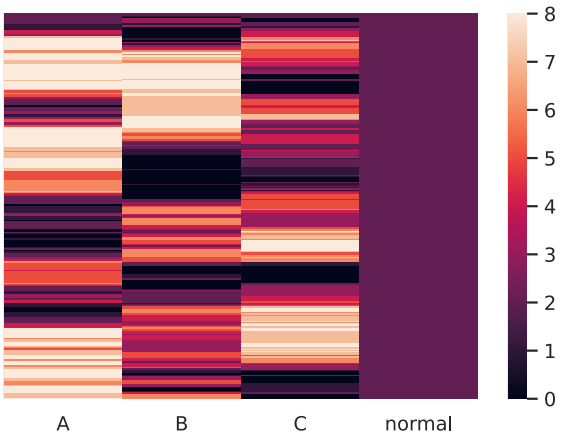


Figure 2.5: Forward simulated CN profiles for three clones (A,B,C) and normal.

2.2.2 Simulating Read Counts

Synthetic read counts were generated using a modified forward sampling (Section 1.5.2) procedure. In this setting, ρ was observed and did not need to be sampled. We computed $\hat{c}_i$, which was a measure of the expected proportion of total reads in bin i. The expected number of reads for a given coverage was determined by multiplying the size of the human genome by the desired coverage and dividing by the average read length. We set the size of the human genome to be 3×10^9 and the average read length to be 135. Then, we sampled from $Multi(\text{reads}, \hat{c})$ using `numpy.multinomial` to get our final dataset. Algorithm 2 describes the read count generation procedure. Figure 2.6 shows scatterplots of forward simulated read counts at various tumour fractions.

2.2.3 Simulating Allelic Counts

Synthetic allelic counts were produced by first sampling from a $Beta(2,2)$ to simulate VAFS for each site/clone pair ($\frac{f'_{lj}}{f_{lj}}$, where $l \in \{1,2,\ldots,250\}$ and $j \in \{1,2,\ldots,k\}$, k being the number of clones). Then, we computed $m_{lj} = \frac{f'_{lj}}{f_{lj}} c_{qj}$, c_{qj} being generated by the process described in Section 2.2.1. We sampled

Algorithm 2: Simulate Read Counts

Input: $\rho \leftarrow [\rho_1, \rho_2, \ldots, \rho_k]$, $c \leftarrow cn\_profiles[n \times k + 1]$

$\bar{c} = (\sum_{j=1}^{k} \rho_j c_{1j}, \sum_{j=1}^{k} \rho_j c_{2j}, \ldots, \sum_{j=1}^{k} \rho_j c_{nj})$

$\hat{c} = \frac{1}{\sum_{l=1}^{n} \bar{c}_l} (\bar{c}_1, \bar{c}_2, \ldots, \bar{c}_n)$

$reads = \frac{cov * 3 \times 10^9}{135}$

$y \sim Multi(reads, \hat{c})$

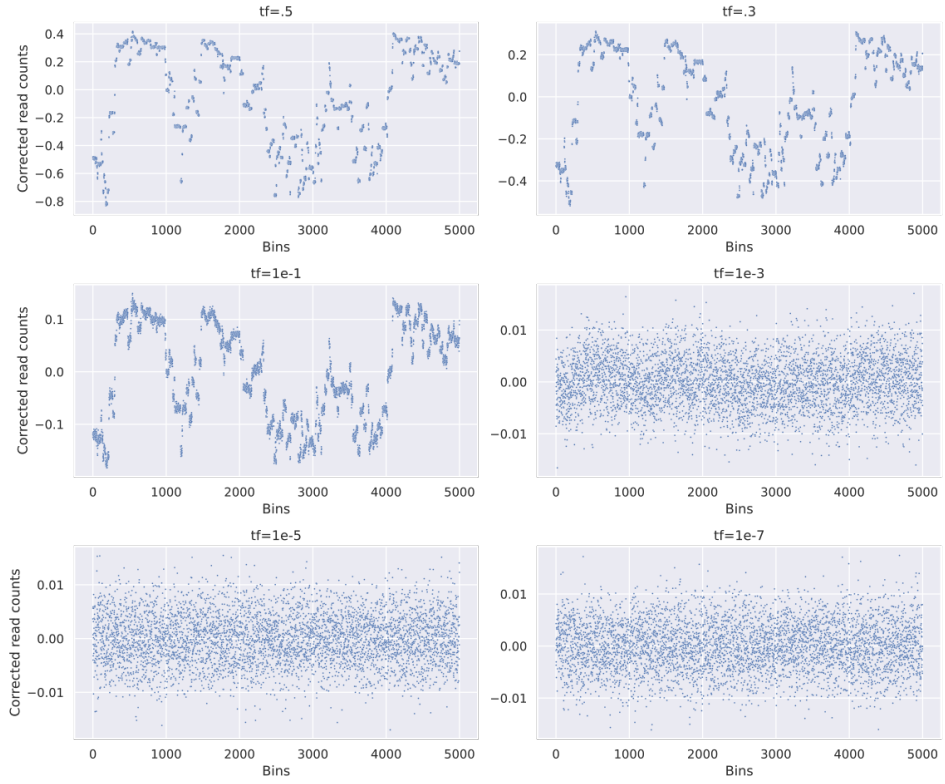


Figure 2.6: Forward simulated read counts.

$d_l \sim Poisson(\text{coverage})$ and $b_l \sim BinLogits(d_l, \xi_l)$, where $\xi_l = \sum_j \rho_j m_{lj}$. Algorithm 3 details the simulation methodology and Equation 2.1 offers a summary of equations and distributions used for this task.

Algorithm 3: Simulate Allelic Counts

Input: $\rho \leftarrow [\rho_1, \rho_2, \ldots, \rho_k]$, $c \leftarrow cn\_profiles[n \times k + 1]$
$s \leftarrow 250$
$m \leftarrow [s \times k]$
$\xi \leftarrow [s \times 1]$
$d \leftarrow [s \times 1]$
$b \leftarrow [s \times 1]$
for $l \leftarrow 0$ **to** $s - 1$ **do**
  for $j \leftarrow 0$ **to** $k - 1$ **do**
    $f \sim Beta(2,2)$
    $q \leftarrow$ bin index containing site l
    $m[l][j] \leftarrow f \times c[q][j]$
  end
  $\xi[l] \leftarrow \sum_{j=1}^{k} \rho_j m_{lj}$
  $d[l] \sim Poisson(cov)$
  $b[l] \sim BinLogits(d[l], \xi[l])$
end

$$
\begin{aligned}
\frac{f'_{lj}}{f_{lj}} &\sim Beta(2,2) \\
m_{lj} &= \frac{f'_{lj}}{f_{lj}} c_{qj} \\
\xi_l &= \sum_j \rho_j m_{lj} \\
d_l &\sim Poisson(cov) \\
b_l &\sim BinLogits(d_l, \xi_l)
\end{aligned}
\tag{2.1}
$$

2.3 Semi-realistic Datasets

2.3.1 Lymphoma Dataset

Single-cell Lymphoma patient data from the Terry Fox Research Institute (TFRI) was obtained using DLP+[36] (Section 1.3.2). Two time points corresponding to Follicular Lymphoma (FL) and Diffuse Large B-Cell Lymphoma (DLBCL) were identified. In our experiments, we did not distinguish between time points. Given the output from DLP+, CN profiles of single-cells were estimated using HMMCopy[34] and clone populations were inferred by cutting the outputted tree from Sitka[52]. Single-cell bam files were merged according to their clone membership using `samtools merge`[11]. To obtain the CN profile of a clone, we took the mean CN value for each bin over all cells assigned to that clone. A total of 6 distinct clone populations (A-F) were distingushed and tissue from a healthy patient was sequenced as a matched normal. Data is available upon request. See Figure 2.7 for a heat map of the CN profiles from the Lymphoma dataset.

2.3.2 Semi-realistic Data Simulation

We downsampled clone-level and matched normal bam files (Section 2.3.1) to imitate CTDNA using `DownsampleSam` from GATK4[12]. For downsampling clone-level bam files, we altered `DownsampleSam`'s `-P` parameter depending on the desired tumour fraction and clonal prevalences. Specifically, we computed `-P` using the following equation,

$$P_j = \frac{r * \rho_j * tf}{n_j} \tag{2.2}$$

where P_j was the value of `-P` in `DownsampleSam` for clone j, r was the target number of reads in the final dataset, ρ_j was the unnormalized clonal prevalence for clone j, tf was the tumour fraction and n_j was the total number of reads in the clone-level bam for clone j. The value of ρ depended on the number of clones in the dataset and is documented in Table A.1. For downsampling the matched

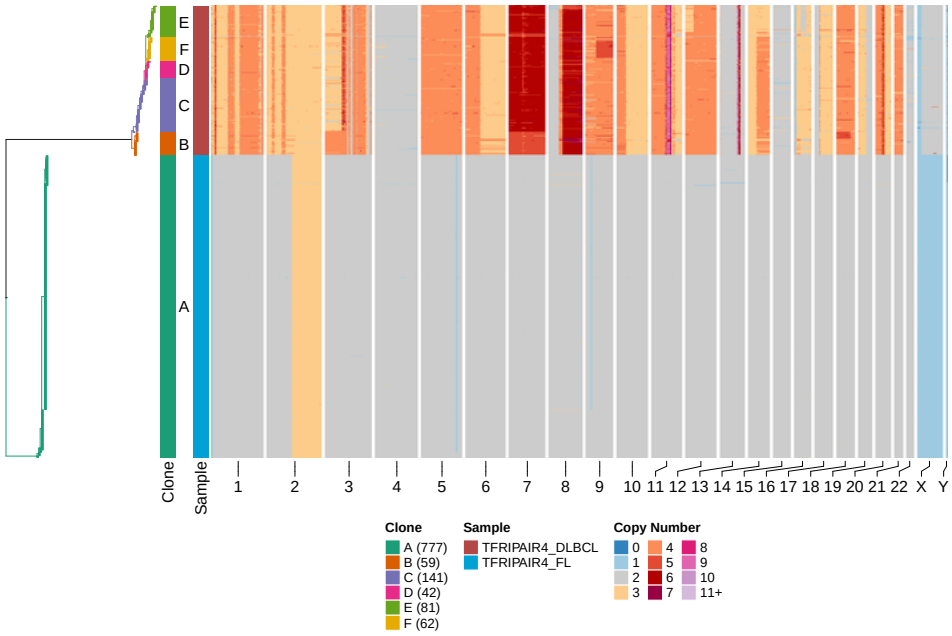


Figure 2.7: Heat map of inferred CN profile for Lymphoma patient data. Each row of the heat map correspond to a single cell. CN values are color coded - red for larger values and blue for smaller. Tree on leftmost side describes the hierarchical clustering process. Colored bars labeled clone and sample delineate the clone and sample a cell is from.

normal bam, we calculated,

$$P_{norm} = \frac{r * (1 - tf)}{n_{norm}} \tag{2.3}$$

where P_{norm} was the value of `-P` in `DownsampleSam` for the matched normal and n_{norm} was the total number of reads in the matched normal bam file. Finally, we merged downsampled clone and normal bams to manufacture an *in-silico* mixture of reads from distinct clonal genomes. We changed the number of clones by including or excluding clone-level bam files during downsampling and merging. We altered tf to obtain semi-realistic datasets with different tumour fractions. Semi-realistic datasets of varying read depths were created by adjusting r in Equation 2.2.

2.3.3 Limitations

Here we note limitations to our semi-realistic data simulation methodology. Nucleosome occupancy and fragmentation patterns in CTDNA have proven to be powerful biomarkers. Nucleosome footprints have effectively inferred cell types contributing CFDNA in cancer [56] and a machine learning model has successfully applied CTDNA fragmentation patterns to predict tissue of origin in cancer [8]. Currently, our semi-realistic data simulation (Section 2.3) ignores nucleosome occupancy and fragmentation patterns, potentially introducing unwanted bias to our semi-realistic datasets.

2.4 Benchmarking

We benchmarked LiquidBayes' performance against two state-of-the-art (SOTA) methods: ichorCNA[1] and MRDetectSNV[69]. We evaluated LiquidBayes' performance on a variety of tumour fractions, read depths and numbers of clones. For each arrangement, we simulated ten datasets and applied each inference method to all datasets. L1 losses were calculated by differencing tumour fraction estimates from the ground truth and then taking the absolute value. L1 losses were divided by tumour fraction and $\log(x+1)$ transformed to compare relative performance between different tumour fraction levels. Finally, we generated plots to visualize and compare performance accuracy across methods. The entire benchmarking pipeline was implemented using Snakemake[42] and is available upon request.

2.4.1 LiquidBayes

We used LiquidBayes v.0.10 in our benchmarking pipeline. The input consisted of an *in silico* mixture of reads simulating CTDNA, GC and mappability wig files generated using `gcCounter` and `mapCounter` from hmmcopy_utils[33], CN profiles of clones, the model type (base or extended), 10000 inference samples, 500 warmup samples and a unique integer to control `numpyro`'s random seed (https://github.com/Roth-Lab/LiquidBayes). LiquidBayes generated samples from the posterior distribution based on the graphical model designated by the model type (Figure 2.1b, Figure 2.2b). To obtain tumour fraction point estimates, we computed the mean over normal fraction estimates and took the complement. Similarly,

we computed the mean over clonal prevalence estimates to get clonal prevalence point estimates.

Variant Calling

Variant calling only pertained to the extended model (Section 2.1.1). We performed somatic variant calling on semi-realistic ctDNA datasets (Section 2.3) and all clone-level pseudobulk genomes (Section 1.3.2) using Strelka2 v.2.9.10[28]. First, we configured the workflow by executing `configureStrelkaSomaticWorkflow.py`. In the configuration step, we used a matched normal and the GRCh37-lite reference genome (https://www.bcgsc.ca/downloads/genomes/9606/hg19/1000genomes/bwa_ind/genome/README.GRCh37-lite). In accordance with best practices, small indel candidates were discovered using Manta[6] with the same sample, matched normal and reference genome. Subsequently, we ran `runWorkflow.py` built during the configuration step. Strelka2 reported all variant predictions in VCF 4.1[10].

2.4.2 ichorCNA

We used ichorCNA[1] v.0.3.2 and followed the supplied Snakefile for proper execution. ichorCNA only required a plasma bam file and did not make use of a matched tissue biopsy. First, we quantified read counts in the plasma bam using hmmcopy_utils `readCounter` [33] with parameters `binSize: 500000`, `qual: 20` and `chrs: 1,2,...,22`. Next, we executed `runIchorCNA.R` with parameters `chrs:1,2,...,22`, `ichorCNA_chrs: c(1:22)`,`ichorCNA_gcWig: gc_hg19_500kb.wig` and `ichorCNA_mapWig: map_hg19_500kb.wig`; `wig` files were located in `inst/extdata/` provided by ichorCNA. Tumour fraction estimates were acquired directly from the output file {id}.params.txt.

2.4.3 MRDetectSNV

No modifications were made to the MRDetect software. MRDetect required two inputs: a plasma bam and a VCF file from the patient-specific tumour biopsy. The VCF file was generated by performing the steps described in Section 2.4.1

on merged clone-level pseudobulk genomes. Tumour fraction was computed using the equation,

$$TF = 1 - (1 - [M - \mu * R]/N)^{1/cov} \tag{2.4}$$

where TF denoted the tumour fraction, M denoted the number of SNVS from the plasma bam, N denoted the number of SNVS in the patient-specific tumour biopsy, R denoted the total number of reads covering the patient-specific tumour biopsy, cov denoted the local coverage in sites with a tumour-specific SNV and μ denoted the noise rate $\frac{\#errors}{\#reads}$. R, M and N were reads_checked, sites_detected and sites_checked, respectively, from MRDetect's output file. μ was obtained by running MRDetect on a matched normal sample and extracting detection_rate. cov was determined using `pysam.depth` (https://github.com/pysam-developers/pysam) with the `-a` flag and computing mean depth on the plasma bam.

Chapter 3

Results

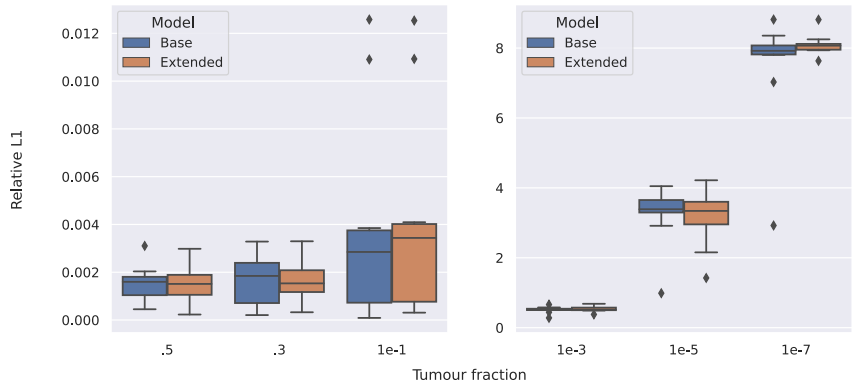
3.1 Synthetic Experiments

We evaluated LiquidBayes' base and extended models by analyzing inference outputs on synthetic datasets with varied tumour fractions, read depths and numbers of clones (Section 2.2). We used the L1 and Relative L1 losses to evaluate accuracy.

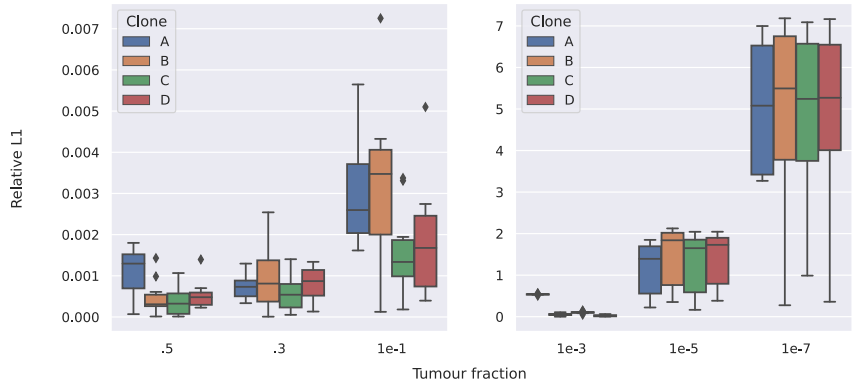
3.1.1 Tumour Fraction

MRD is a major cause of relapse in post-treatment cancer patients. Tracking tumour burden in patients can aid in MRD discovery and improve prognosis. Although liquid biopsies facilitate serial sampling, which is vital for relapse surveillance, the tumour burden in a blood draw can fall to proportions as low as 1e-5. Accordingly, statistical methods have been developed to estimate tumour burden from liquid biopsies, but struggled when tumour proportions fell below 1e-5.

We evaluated LiquidBayes at successively smaller tumour fractions (tf={.5, .3, 1e-1, 1e-3, 1e-5, 1e-7}). In general, we observed that the error increased as the tumour fraction decreased (Figure 3.1a, Table 3.1). Both models performed comparably and gave accurate predictions at tumour fractions greater than 1e-3. LiquidBayes distinguished itself from other models by inferring the prevalence of individual clonal populations in addition to the overall tumour fraction. In our experiments, LiquidBayes returned accurate clonal prevalence estimates for tumour



(a) Boxplots of logged relative-error of tumour fraction estimates for six tumour fraction levels for the base and extended models. Synthetic datasets had 5 clones and a 1x read depth.



(b) Boxplots of logged relative-error of tumour and clone fraction estimates for six tumour fraction levels using the base model. Simulated datasets had a read depth of 1x and four clones to demonstrate LiquidBayes' capacity to determine clonal prevalences.

Figure 3.1: Boxplots summarizing results for tumour fraction synthetic experiments. Ten datasets were created for each tumour fraction setting.

fractions greater than 1e-5 (Figure 3.1b, Table 3.2). See Figure A.1 and Table A.2 for results on the extended model.

We plotted MCMC samples for a single replicate at each tumour fraction (Figure 3.2). For both models, the HDIS for tumour fractions greater than 1e-3 contained the true value, whereas the HDIS at lower tumour fractions did not. The

Model	Tumour fraction	L1	Relative L1
Base	.5	0.003669	0.007296
	.3	0.002758	0.009144
	1e-1	0.002371	0.023308
	1e-3	0.000819	0.592830
	1e-5	0.000074	1.377229
	1e-7	0.000145	6.114356
Extended	.5	0.003706	0.007363
	.3	0.002179	0.007228
	1e-1	0.002185	0.021505
	1e-3	0.000894	0.637834
	1e-5	0.000067	1.363940
	1e-7	0.000098	5.571932

Table 3.1: Average L1 and Relative L1 values across ten replicates for synthetic experiments on the base and extended models at six tumour fraction levels. Relative L1=log(L1 / tumour fraction + 1).

posterior plots for tumour fractions .5, .3 and 1e-1 were unimodal, whereas the posterior plots for tumour fractions 1e-3, 1e-5 and 1e-7 were multimodal.

We visualized the distribution of HDI widths across all replicates at each tumour fraction level (Figure 3.3a). We observed that the width decreased along with tumour fraction. Moreover, there was no perceptible change in HDI widths at tumour fractions under 1e-3. Figure 3.3b illustrates the Bayesian coverage (Section 1.5.4) at each tumour fraction. LiquidBayes had high coverages at large tumour fractions (.5, .3, 1e-1) and zero coverages at small tumour fractions (1e-3, 1e-5, 1e-7).

3.1.2 Read Depth

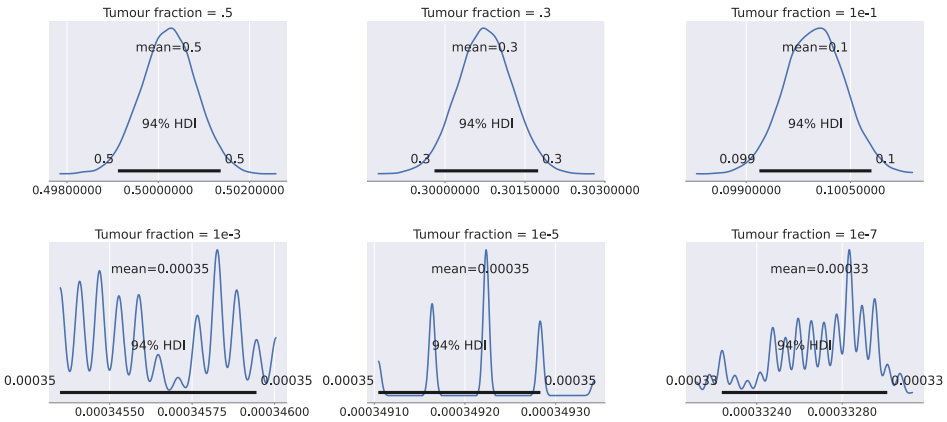
Read depth is defined as the number of times individual bases have been sequenced. Oftentimes, it is helpful to examine the average read depth, as it provides a rough measure of the signal-to-noise ratio. Herein, we use read depth to reference average read depth. Naturally, higher read depths are preferable, but also more costly. For a statistical model, it is advantageous to ascertain read depth's influence on accuracy and to identify a range of read depths in which reliable results can be obtained.

Tumour fraction	Clone	True proportion	L1	Relative L1
.5	A	0.35	0.000569	0.001138
	B	0.1	0.000233	0.000466
	C	0.025	0.000199	0.000398
	D	0.025	0.000267	0.000534
.3	A	0.21	0.000220	0.000731
	B	0.06	0.000289	0.000964
	C	0.015	0.000174	0.000580
	D	0.015	0.000241	0.000804
1e-1	A	7e-2	0.000309	0.003088
	B	2e-2	0.000317	0.003162
	C	5e-3	0.000155	0.001546
	D	5e-3	0.000186	0.001856
1e-3	A	7e-4	0.000713	0.538040
	B	2e-4	0.000052	0.050590
	C	5e-5	0.000108	0.102501
	D	5e-5	0.000028	0.027056
1e-5	A	7e-6	0.000028	1.177543
	B	2e-6	0.000043	1.471758
	C	5e-7	0.000036	1.313933
	D	5e-7	0.000039	1.411559
1e-7	A	7e-6	0.000038	5.030229
	B	2e-8	0.000047	4.966509
	C	5e-9	0.000040	4.900767
	D	5e-9	0.000041	4.940837

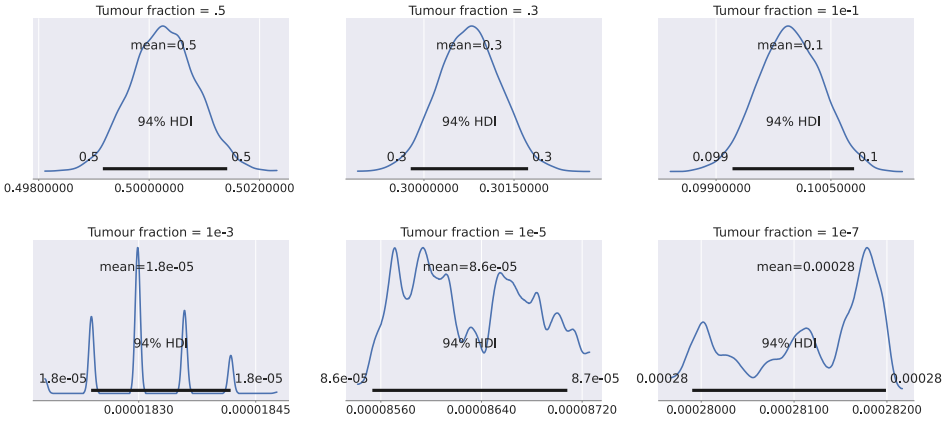
Table 3.2: Average L1 and Relative L1 values for clone fraction estimates across ten replicates for synthetic experiments on the base model at six tumour fraction levels. Relative L1=log(L1 / tumour fraction + 1).

In doing so, one can optimize between cost and accuracy depending on resource accessibility.

We analyzed LiquidBayes' performance on synthetic datasets of diverse read depths. Overall, estimates were better and less variable at higher read depths (Figure 3.4, Table 3.3). We also tested LiquidBayes at sequentially smaller tumour fractions to determine when read depth no longer affected accuracy. At tumour fractions 1e-1 and 1e-2, read depths larger than 1e-3x produced acceptable esti-



(a) Base model.



(b) Extended model.

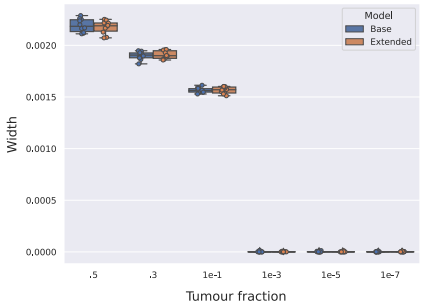
Figure 3.2: Synthetic experiments posterior plots from both models at six tumour fraction levels. Dark black lines indicate the HDI.

mates. For tumour fraction = 5e-3, we observed subpar solutions at read depths smaller than .5x, with estimates being over 40% off of the ground truth. Eventually, at tumour fraction = 1e-3, we observed that read depth had no effect on accuracy. Numeric results for the extended model can be found in Table A.3.

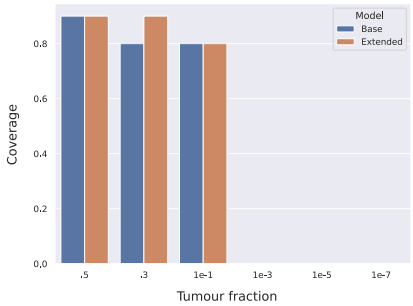
We plotted MCMC samples from experiments on each read depth for tumour fractions 5e-3 and 1e-3 (Figure 3.5). We chose these tumour fractions because they marked a shift as to read depth's impact on accuracy. At tumour fraction = 5e-3, posterior plots for all read depths were unimodal (Figure 3.5a). Moreover,

Tumour fraction	Read depth	L1	Relative L1
1e-1	100	0.000343	0.003425
	10	0.000380	0.003789
	1	0.000673	0.006689
	.5	0.000830	0.008259
	1e-1	0.003093	0.030304
	1e-2	0.005928	0.057149
	1e-3	0.012274	0.112380
1e-2	100	0.000035	0.003498
	10	0.000204	0.020163
	1	0.001493	0.122506
	.5	0.000962	0.089021
	1e-1	0.001305	0.120614
	1e-2	0.003062	0.253836
	1e-3	0.010376	0.688400
5e-3	100	0.000047	0.009416
	10	0.000105	0.020579
	1	0.000695	0.126902
	.5	0.000688	0.123607
	1e-1	0.002202	0.339780
	1e-2	0.004443	0.615581
	1e-3	0.014065	1.288322
1e-3	100	0.000806	0.589036
	10	0.000279	0.236296
	1	0.000909	0.644174
	.5	0.000866	0.621109
	1e-1	0.002063	1.047924
	1e-2	0.000815	0.593020
	1e-3	0.000749	0.557197

Table 3.3: Average L1 and Relative L1 losses across ten replicates for synthetic experiments on the base model at seven read depths and four tumour fraction levels. Relative L1=log(L1 / tumour fraction + 1)



(a) Distribution of 94% HDI widths across ten replicates for the base and extended models at six tumour fractions. A strip plot is superimposed for easily identifying model types.



(b) Bar plot depicting the Bayesian coverage across ten replicates at six tumour fractions.

Figure 3.3: Posterior statistics for synthetic experiments on tumour fractions. **(a)** HDI widths. **(b)** Bayesian coverage.

the HDI for all read depths except 1e-2x and 1x contained the true tumour fraction. At tumour fraction = 1e-3, we saw multimodality in the posterior plots at all read depths except 1e-1x and 10x (Figure 3.5b). Furthermore, only the HDIS for read depths 1e-1x and 10x contained the true tumour fraction. MCMC plots from the extended model can be found in Figure A.2.

The width of the HDI decreased as the read depth increased (Figure A.3a). Read depth and coverage did not exhibit any clear patterns (Figure A.3b). We observed a downward trend in coverage across all read depths as tumour fraction decreased.

3.1.3 Number of Clones

Tumour heterogeneity is coincident with multiple clonal populations. Moreover, the amount of clones in the tumour can fluctuate over time. Therefore, it is valuable to produce accurate estimates notwithstanding the number of clones.

We assessed how the number of clones affected LiquidBayes' performance using synthetic datasets containing between two to six clones. For the most part, error declined as the number of clones increased (Figure 3.6, Table 3.4). This trend was more pronounced at tumour fraction = 1e-1, but slightly ambiguous for tumour

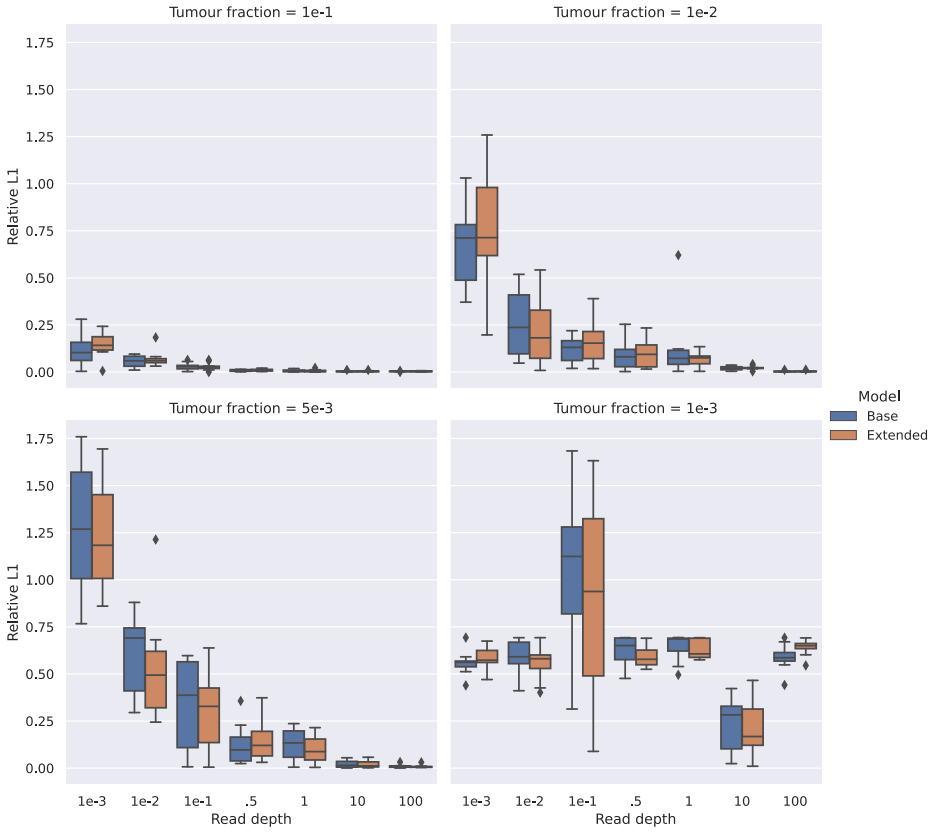
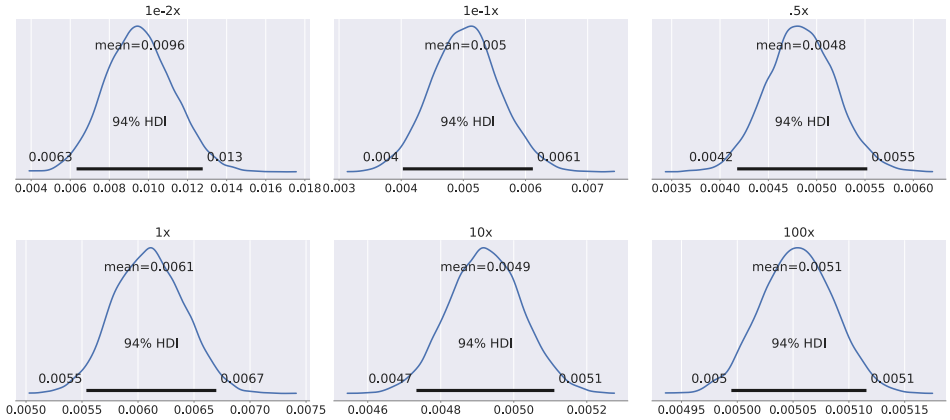


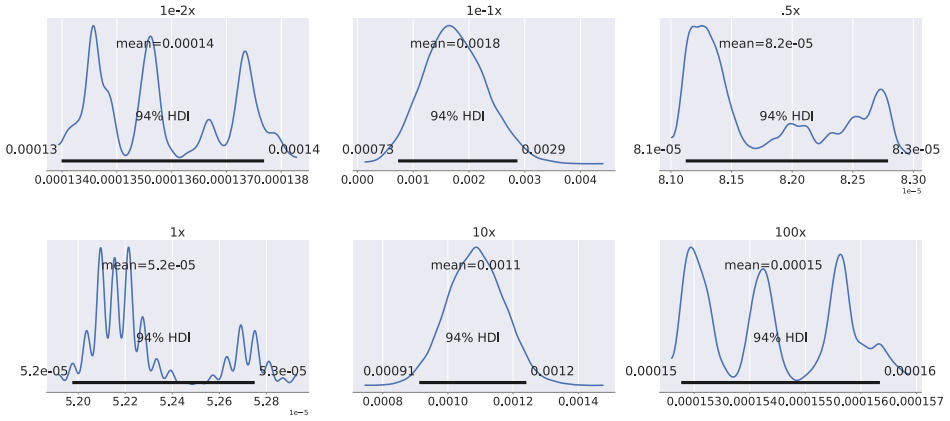
Figure 3.4: Boxplots of Relative L1 in synthetic experiments for seven read depths at four tumour fraction levels. Read depth was inversely related to error. The variance of Relative L1 losses was more volatile at read depths less than .5x than at read depths greater than .5x. Simulated datasets included 3 clones and 10 replicates.

fraction = .5. Furthermore, the variability of estimates decreased as the number of clones increased. We did not observe any significant difference in performance between the base and extended models. L1 and Relative L1 values for the extended model can be found in Table A.4.

Posterior plots of MCMC samples from the base model (Figure 3.7) and extended model (Figure A.4) were generated for these experiments. Posterior means



(a) Tumour fraction = 5e-3.



(b) Tumour fraction = 1e-3.

Figure 3.5: Posterior plots from synthetic experiments using the base model at six read depths for tumour fractions 5e-3 and 1e-3. Bold black lines indicate the HDI.

were extremely close to the true tumour fraction for all clone quantities and all posterior plots were unimodal.

There was no relationship between the number of clones and the HDI width (Figure 3.8a). Though, HDI widths were smaller and more concentrated at tumour fraction = 1e-1. The coverage was generally higher for larger numbers of clones at both tumour fraction settings (Figure 3.8b). Surprisingly, the coverage was much larger at tumour fraction = 1e-1 at three clones.

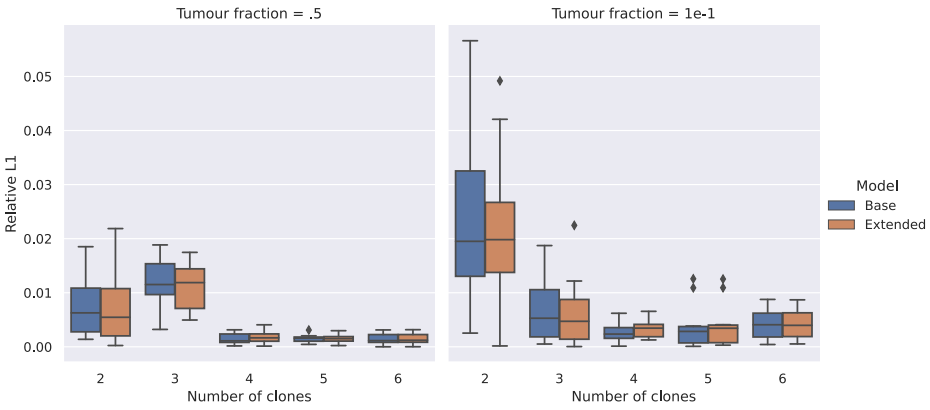
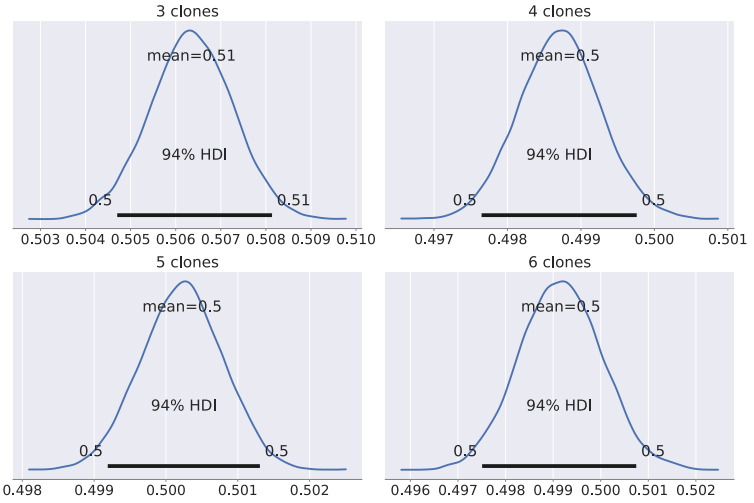


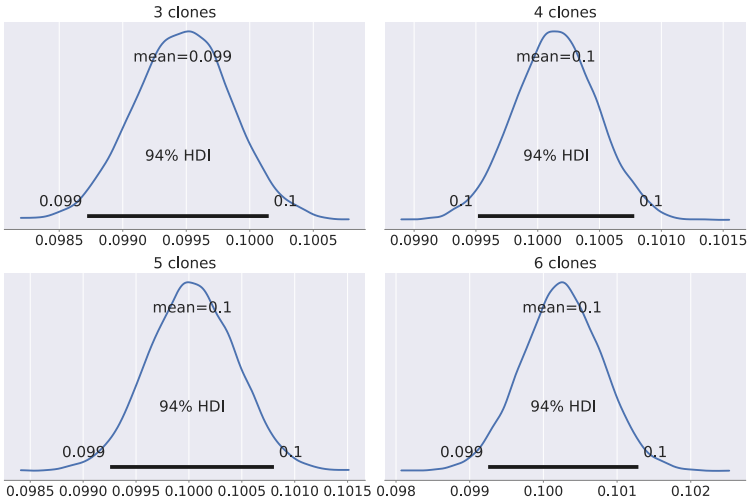
Figure 3.6: Boxplots of Relative L1 values from synthetic experiments for two to six clones at two tumour fractions. Simulated datasets had an average read depth of 1x and contained ten replicates.

Tumour fraction	Number of clones	L1	Relative L1
.5	2	0.003669	0.007296
	3	0.005946	0.011811
	4	0.000761	0.001521
	5	0.000746	0.001490
	6	0.000705	0.001409
1e-1	2	0.002371	0.023308
	3	0.000673	0.006689
	4	0.000288	0.002877
	5	0.000392	0.003901
	6	0.000405	0.004034

Table 3.4: Average L1 and Relative L1 values across ten replicates for synthetic experiments on the base model at five different numbers of clones and two tumour fraction levels. Relative L1=log(L1 / tumour fraction + 1)

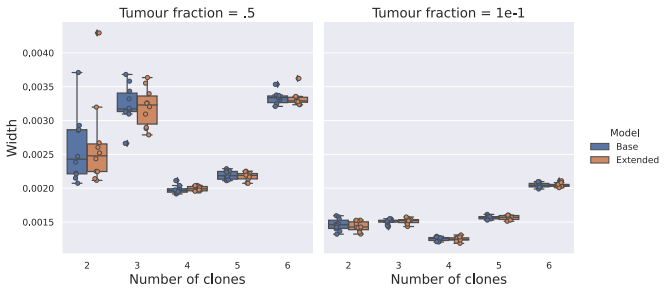


(a) Tumour fraction = .5.

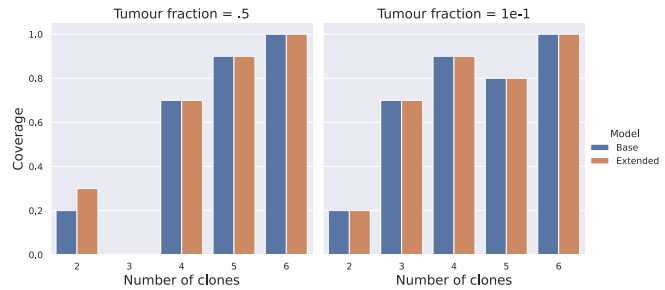


(b) Tumour fraction = 1e-1.

Figure 3.7: Posterior plots from synthetic experiments using the base model for three to six clones at two tumour fractions. Bold black lines indicate the HDI.



(a) Distributions of 94% HDI widths across ten replicates for the base model for two to six clones at two tumour fractions. A strip plot is superimposed for easily identifying model types.



(b) Bar plots depicting the Bayesian coverage across ten replicates for two to six clones at two tumour fractions.

Figure 3.8: Posterior distribution statistics for synthetic experiments on different numbers of clones. **(a)** HDI widths. **(b)** Bayesian coverage.

3.1.4 Missing Clones

Novel clones can emerge over time according to clonal evolution. Consequently, these novel clones may not be covered by the set of CN profiles derived from a prior tissue biopsy. We recognized this to be a potential shortcoming of LiquidBayes, so we tested its robustness to missing clones.

We mimicked instances where the prior tissue biopsy did not fully characterize all clone populations by removing the CN profile of the smallest or largest clone proportion-wise prior to inference (Section 2.2.1). Table 3.5 documented the proportion values of removed clones. The error increased moderately when we removed the smallest clone, but increased considerably when we removed the largest clone (Figure 3.9). When there were four or more clones, removing the smallest clone slightly affected performance. We saw a big difference removing the largest

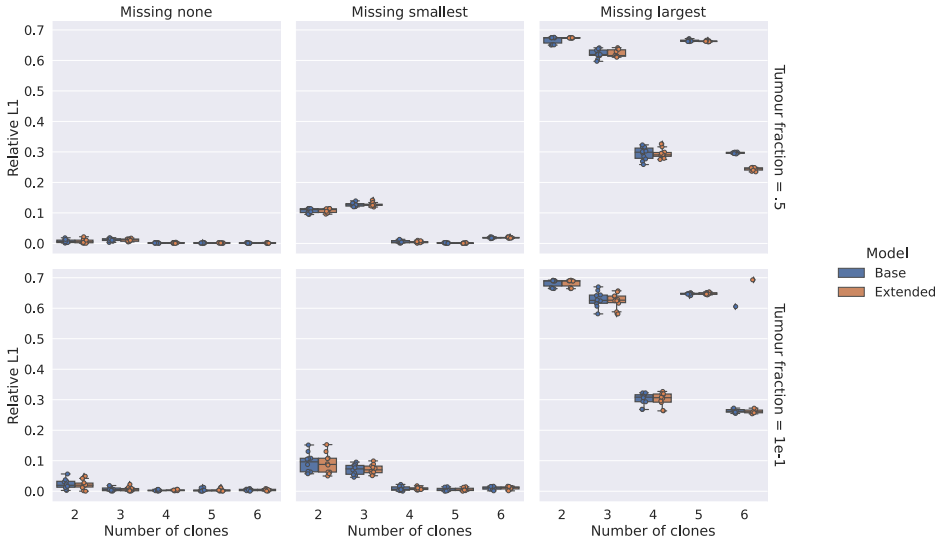


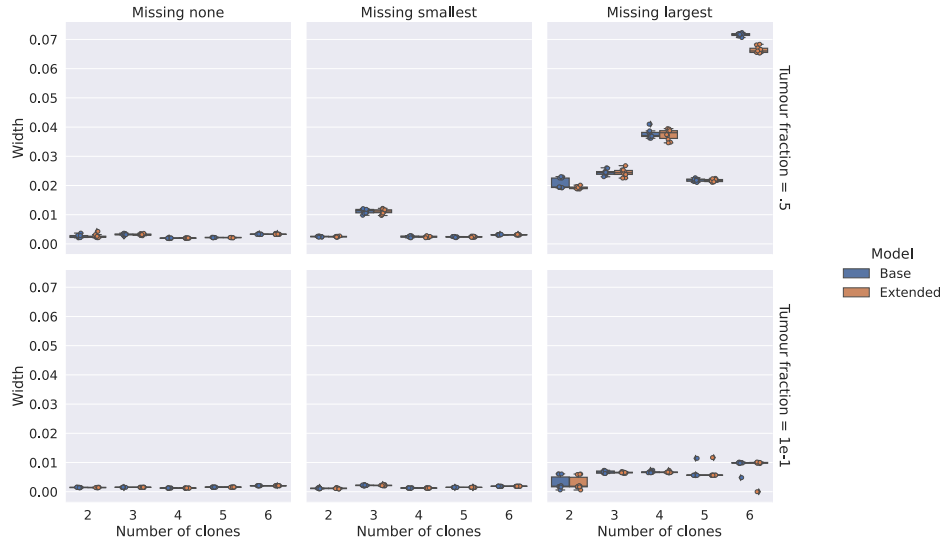
Figure 3.9: Boxplots with superimposed strip plots of Relatve L1 values for two to six clones displaying the effect of removing the smallest or largest clone. Table 3.5 documents the proportions of the removed clones. Synthetic datasets had a read depth of 1x.

clone when only having a small number of clones. See Figure A.5 and Figure A.6 for MCMC plots from the base model. See Figure A.7 and Figure A.8 for MCMC plots from the extended model.

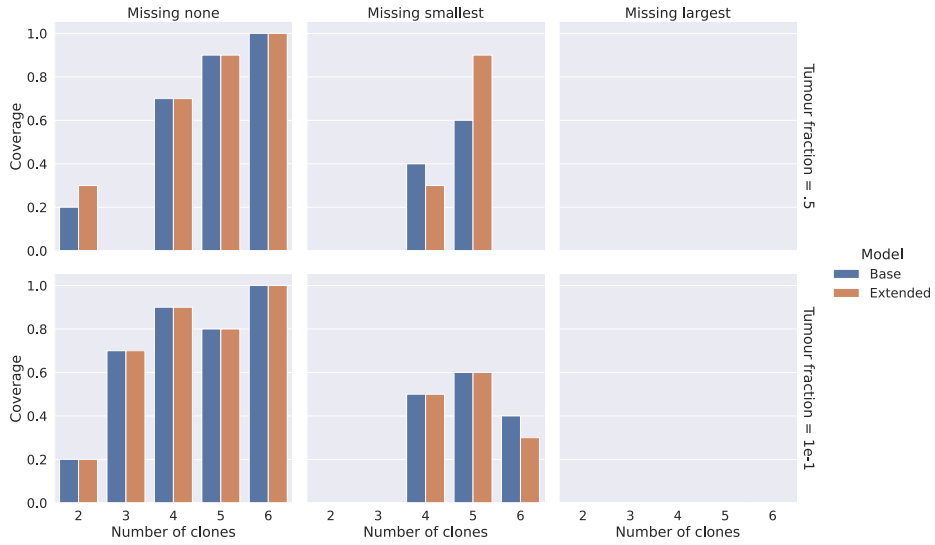
There was no significant change in the distribution of HDI widths from omitting the smallest clone (Figure 3.10a). In contrast, the distribution of HDI widths increased considerably when the largest clone was removed, especially at tumour fraction = .5. Figure 3.10b showed that removing a clone caused the coverage to decrease. When the smallest clone was removed, the coverage was greater than zero at some quantities of clones. However, when the largest clone was removed, the coverage was zero throughout.

3.2 Semi-realistic Experiments

We executed experiments on idealized mixtures of clonal populations (semi-realistic datasets) from DLP+ on single-cells from a Lymphoma patient (Sec-



(a) Distribution of 94% HDI widths across ten replicates for two to six clones (includes the missing clone) at tf=.5 and 1e-1. A strip plot is superimposed for easily identifying model types.



(b) Bar plot depicting the Bayesian coverage across ten replicates for two to six clones (includes the missing clone) at two tumour fractions.

Figure 3.10: Posterior distribution statistics for synthetic experiments on the effects of a missing clone. **(a)** HDI widths **(b)** Bayesian coverage

Number of Clones	Tumour fraction	Smallest	Largest
2	.5	0.049050	0.450950
	1e-1	0.009810	0.090190
3	.5	0.058240	0.369840
	1e-1	0.011650	0.073970
4	.5	0.003410	0.380530
	1e-1	0.000680	0.076110
5	.5	0.004210	0.421630
	1e-1	0.000840	0.084330
6	.5	0.005260	0.339960
	1e-1	0.001050	0.067990

Table 3.5: Proportions of the smallest and largest clones that were removed in synthetic experiments at different numbers of clones and tumour fractions.

tion 2.3). These experiments demonstrated how LiquidBayes might behave in real-world contexts and how it measured up to the current SOTA methods (ichorCNA and MRDetectSNV). Similar to Section 3.1, we used the L1 loss and Relative L1 loss to measure the quality of estimates.

3.2.1 Tumour Fraction

We applied LiquidBayes, ichorCNA and MRDetectSNV on semi-realistic datasets with varied tumour fractions. LiquidBayes' base and extended models significantly outperformed both ichorCNA and MRDetect at all tumour fractions (Figure 3.11, Table 3.6). Furthermore, the gap in accuracy between LiquidBayes and the other two models grew larger as the tumour fraction decreased. Also, LiquidBayes' estimates were stable and clustered around the medians.

To visualize uncertainty, we plotted MCMC samples from LiquidBayes' base and extended models for a single replicate at each tumour fraction (Figure 3.12). We observed short HDI widths, with none containing the true tumour fraction. All posterior distributions were unimodal.

There was a steady decline in HDI widths as the tumour fraction decreased (Figure 3.13a). There were also a few outliers at tumour fractions .3 and 1e-1. We

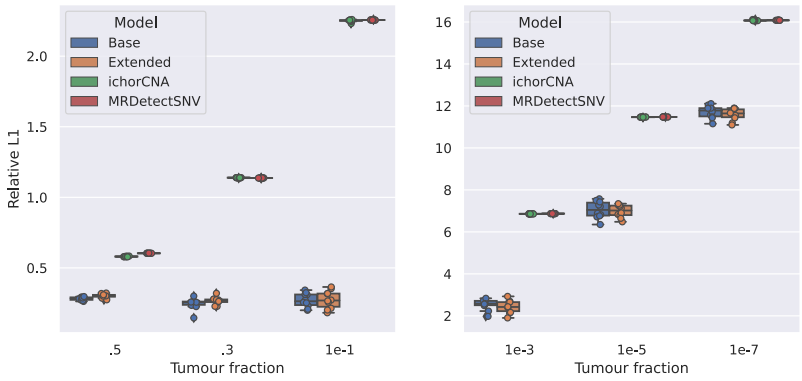


Figure 3.11: Summary of results from LiquidBayes, ichorCNA and MRDetectSNV for tumour fraction experiments on semi-realistic datasets. LiquidBayes' base and extended models were comparable in accuracy. Datasets had three clones and a read depth of 1x.

observed low coverages across all tumour fractions (Figure 3.13b), with tumour fractions 1e-1 and 1e-3 being the only ones with a coverage greater than zero.

We also explored LiquidBayes' ability to estimate individual clone proportions. Relative L1 values increased as tumour fraction decreased (Figure 3.14). Relative L1 values associated with clone A was largest among the three clones. This was most likely because clone A had the highest proportion. For tumour fractions above 1e-3, the L1 was relatively small relative to the true clone proportion (Table 3.7). Boxplots and L1 values for the extended model are in Figure A.9 and Table A.5, respectively.

3.2.2 Read Depth

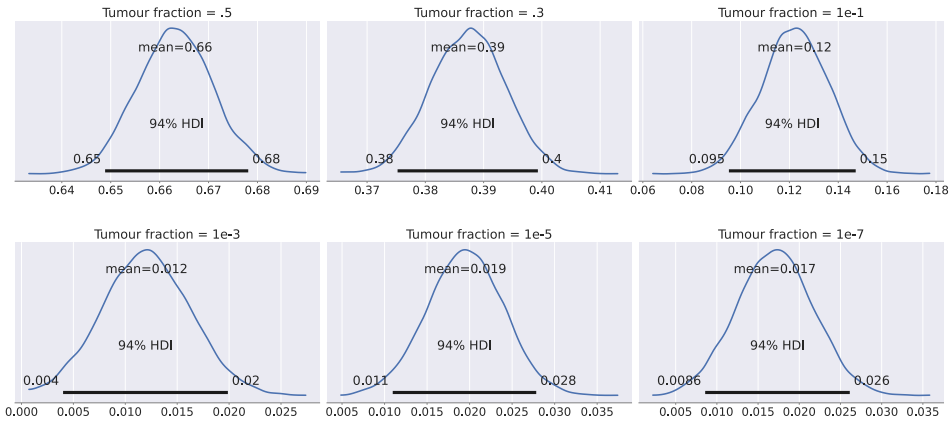
We evaluated read depth's impact on each model. At a read depth of 1e-3x, all models performed comparably. However, at read depths above 1e-3x, LiquidBayes gave better results than the other models (Figure 3.15). Moreover, LiquidBayes' estimates improved with higher read depths. In contrast, both ichorCNA and MRDetectSNV did not benefit from higher read depths. At read depth 1x, LiquidBayes' base and extended models produced estimates with L1 values below 1e-1 at all

Model	Tumour fraction	L1	Relative L1
Base	.5	0.163208	0.282422
	.3	0.083843	0.245732
	1e-1	0.031260	0.270887
	1e-3	0.012071	2.542357
	1e-5	0.012392	7.052994
	1e-7	0.012726	11.715263
Extended	.5	0.175845	0.301263
	.3	0.091578	0.266083
	1e-1	0.031857	0.274746
	1e-3	0.010914	2.435946
	1e-5	0.011260	6.988702
	1e-7	0.011214	11.593793
ichorCNA	.5	0.393600	0.580649
	.3	0.636984	1.138883
	1e-1	0.849707	2.250956
	1e-3	0.951829	6.859417
	1e-5	0.955873	11.467784
	1e-7	0.958093	16.075162
MRDetectSNV	.5	0.414455	0.603720
	.3	0.634819	1.136569
	1e-1	0.854658	2.256182
	1e-3	0.962231	6.870293
	1e-5	0.963455	11.475706
	1e-7	0.963797	16.081221

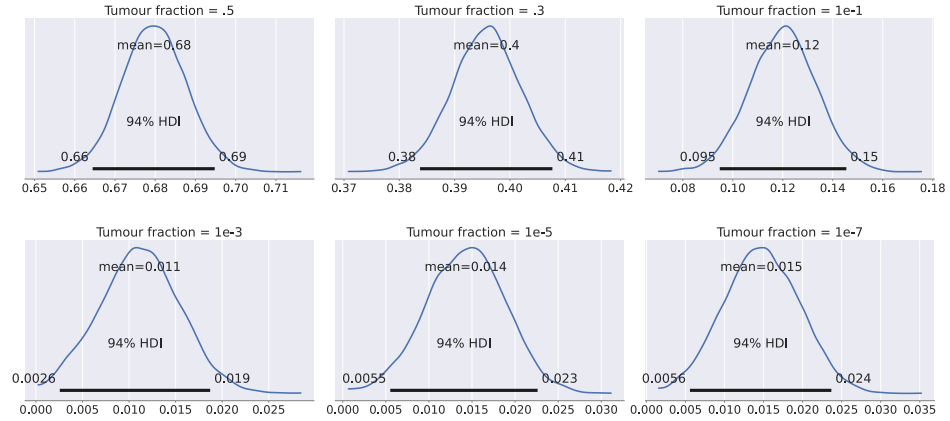
Table 3.6: Average L1 and Relative L1 values across ten replicates for semi-realistic experiments on LiquidBayes' base and extended models, ichorCNA and MRDetectSNV at six tumour fractions. Relative L1=log(L1 / tumour fraction + 1).

tumour fraction/read depth combinations, whereas L1 values for ichorCNA and MRDetect were much higher (Table 3.8). Importantly, we noted that MRDetect-SNV had low L1 values when read depth was 1e-3x. Upon inspection, we saw that MRDetectSNV estimated zero tumour burden in these cases, invalidating those results.

We noted that LiquidBayes yielded fluctuating estimates at a read depth of 1e-1x for tumour fractions 1e-3 and 1e-2. Therefore, we chose to plot the MCMC



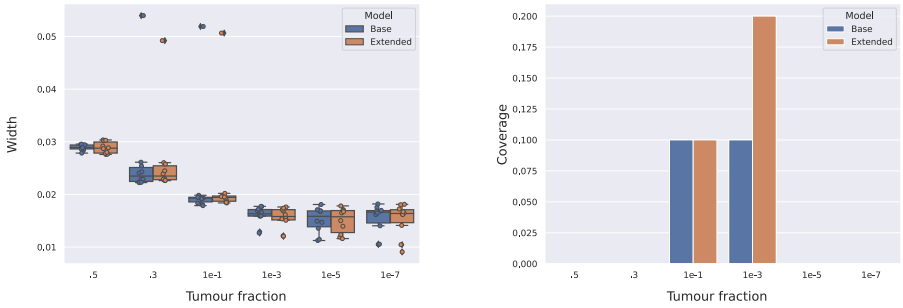
(a) Base model.



(b) Extended model.

Figure 3.12: Semi-realistic experiments posterior plots from both models at six tumour fraction levels. Dark black lines indicate the HDI.

samples of the runs that gave the best and worst estimates for these cases (Figure 3.16). Naturally, we saw a large discrepancy in posterior means between the replicates with the best and worst estimates. Interestingly, we did not see any improvement between replicates at other read depths. In the end, no HDI contained the true tumour fraction value. See Figure A.10a and Figure A.10b for summaries of the posterior distributions.



(a) Distribution of 94% HDI widths across ten replicates for the base and extended models at six tumour fractions. Superimposed strip plots assist in delineating model types.

(b) Bar plot depicting the Bayesian coverage across ten replicates at six tumour fractions.

Figure 3.13: Posterior distribution statistics for semi-realistic experiments on tumour fraction. **(a)** HDI widths **(b)** Bayesian coverage

3.2.3 Number of Clones

To determine how the amount of clones affected accuracy, we executed each model on datasets with different quantities of clones. All models were robust to the number of clones, in that differing clone quantities did not diminish accuracy (Figure 3.17). LiquidBayes produced slightly better estimates at higher numbers of clones, whereas the number of clones did not affect ichorCNA (except at tumour fraction=.5 with five clones) or MRDetect's outputs. Corresponding L1 and Relative L1 values are documented in Table 3.9.

Figure A.11 displays MCMC plots from the base and extended models. The posterior means were stable and somewhat close to the ground truth for all numbers of clones at both tumour fractions. All distributions were unimodal and symmetric.

We saw a slight increase in HDI widths for larger clone quantities (Figure A.12a). This differed from our synthetic experiments (Section 3.1.3) where we observed HDI widths decrease for larger clone quantities. We hypothesized this was an artifact of our CN simulation process which generated independent CN profiles (Section 2.2.1). Overall, we observed zero coverage except at tumour fraction = 1e-1 and 3 clones (Figure A.12b).

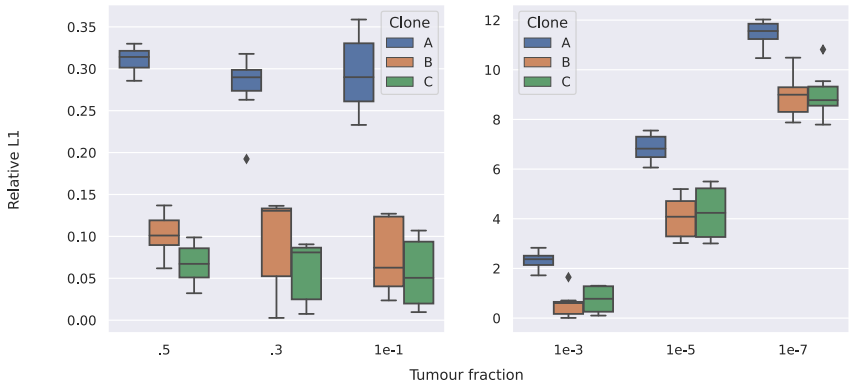


Figure 3.14: Boxplots of Relative L1 values of tumour and clone fraction estimates on semi-realistic datasets for six tumour fraction levels using the base model. Datasets had three clones A,B and C at proportions .8, .15 and .15, respectively and a read depth of 1x.

3.2.4 Missing Clone

We sought to test LiquidBayes' behavior when a clone was missing in the tissue biopsy in the semi-realistic setting. Our design was identical to Section 3.1.4 for removing clones. LiquidBayes outperformed all other methods even when the smallest clone was removed, at both tumour fractions. The only configuration where LiquidBayes' accuracy was comparable to other models was at tumour fraction = .5, with the largest clone removed. At tumour fraction = 1e-1, LiquidBayes gave better estimates than the other models even when the largest clone was removed. Figure A.13 plots posterior statistics from our missing clone experiments.

Clone	Tumour fraction	True proportion	L1	Relative L1
A	.5	0.4	0.182716	0.311380
	.3	0.24	0.097564	0.281023
	1e-1	0.08	0.034300	0.293953
	1e-3	0.0008	0.009737	2.318444
	1e-5	8e-06	0.010570	6.851631
	1e-7	8e-08	0.010458	11.440506
B	.5	0.075	0.053920	0.102116
	.3	0.045	0.030293	0.094941
	1e-1	0.015	0.008047	0.076537
	1e-3	0.00015	0.000928	0.536594
	1e-5	1.5e-06	0.000754	4.032632
	1e-7	1.5e-08	0.001127	8.975954
C	.5	0.025	0.034412	0.066291
	.3	0.015	0.018288	0.058606
	1e-1	0.005	0.005884	0.056475
	1e-3	5e-05	0.001408	0.760778
	1e-5	5e-07	0.001067	4.240922
	1e-7	5e-09	0.001141	8.933280

Table 3.7: Mean L1 and Relative L1 values for clone proportion estimates across ten replicates for semi-realistic experiments on LiquidBayes' base model at six tumour fractions with three clones and a read depth of 1x. Ground truth proportions for clones at each tumour fraction were recorded under the True proportion column. Relative L1=log(L1 / tumour fraction + 1).

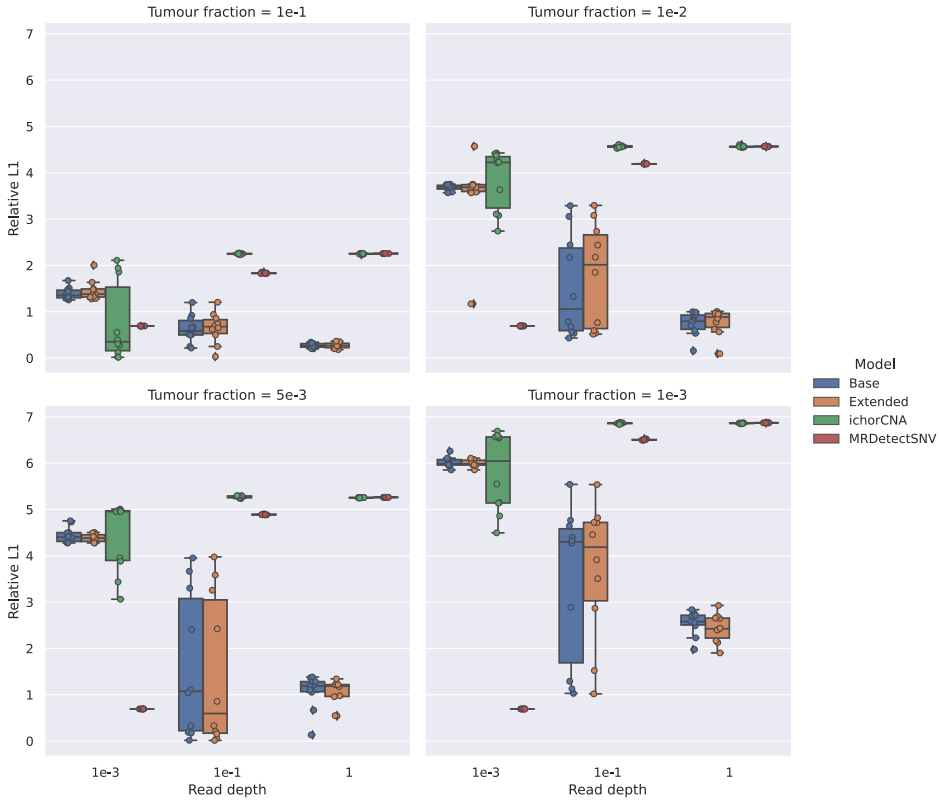
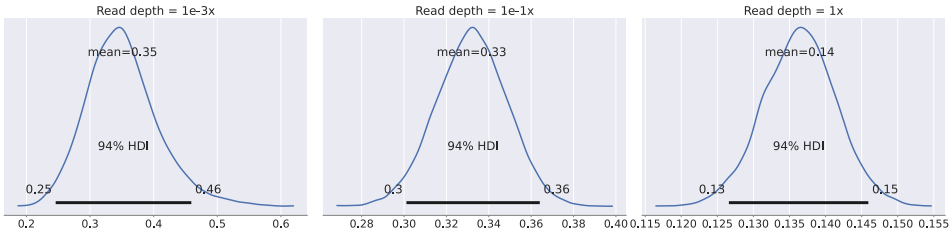


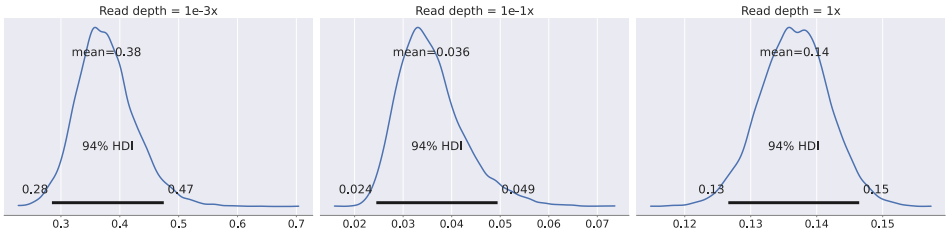
Figure 3.15: Boxplots of Relative L1 values for three read depths at four tumour fraction levels in semi-realistic experiments. Datasets included three clones and ten replicates. Superimposed strip plots assist in delineating model types.

Model	Tumour fraction	Read depth	L1	Relative L1
Base	1e-1	1e-3	0.306816	1.395203
		1	0.031260	0.270887
	1e-2	1e-3	0.387735	3.681347
		1	0.011513	0.738639
	1e-3	1e-3	0.412978	6.019917
		1	0.012071	2.542357
	5e-3	1e-3	0.428585	4.448511
		1	0.010337	1.065394
Extended	1e-1	1e-3	0.340614	1.458136
		1	0.031857	0.274746
	1e-2	1e-3	0.406051	3.513826
		1	0.012345	0.772748
	1e-3	1e-3	0.402231	5.996931
		1	0.010914	2.435946
	5e-3	1e-3	0.396920	4.383550
		1	0.009659	1.041936
ichorCNA	1e-1	1e-3	0.207100	0.760772
		1	0.849707	2.250956
	1e-2	1e-3	0.540980	3.849103
		1	0.951153	4.565434
	1e-3	1e-3	0.445450	5.814030
		1	0.951829	6.859417
	5e-3	1e-3	0.504780	4.416835
		1	0.951555	5.253893
MRDetectSNV	1e-1	1e-3	0.100000	0.693147
		1	0.854824	2.256356
	1e-2	1e-3	0.010000	0.693147
		1	0.952476	4.566921
	1e-3	1e-3	0.001000	0.693147
		1	0.962039	6.870093
	5e-3	1e-3	0.005000	0.693147
		1	0.958327	5.260955

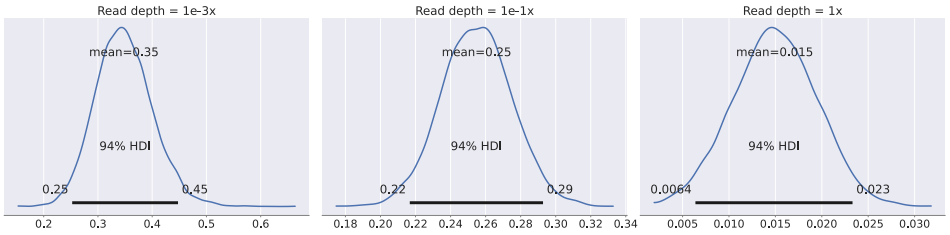
Table 3.8: Mean L1 and Relative L1 values across ten replicates for semi-realistic experiments on LiquidBayes' base and extended models, ichorCNA and MRDetectSNV at two read depths and four tumour fractions. Relative L1=log(L1 / tumour fraction + 1)



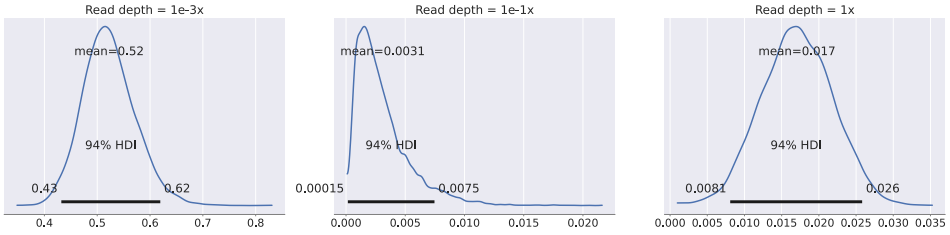
(a) Tumour fraction = 1e-1; replicate with the worst estimate.



(b) Tumour fraction = 1e-1; replicate with the best estimate.



(c) Tumour fraction = 1e-3; replicate with the worst estimate.



(d) Tumour fraction = 1e-3; replicate with the best estimate.

Figure 3.16: Posterior plots from semi-realistic experiments of the best and worst replicates with the 94% HDI from the base model at three read depths for tumour fractions 1e-1 (Figure 3.16a, Figure 3.16b) and 1e-3 (Figure 3.16c, Figure 3.16d). Bold black lines indicate the HDI.

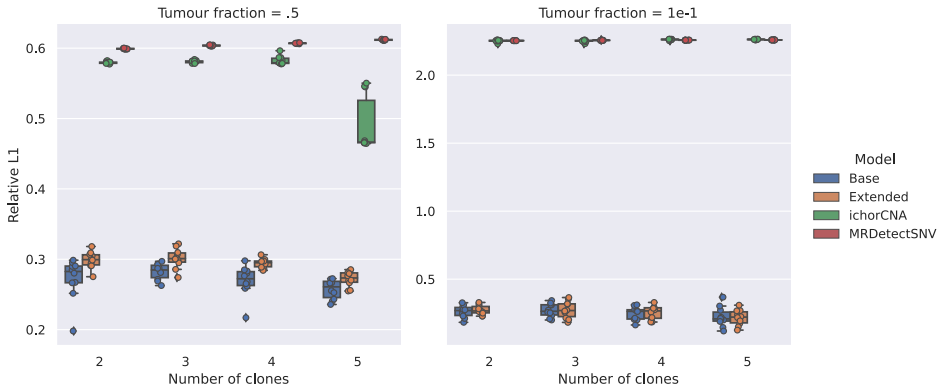


Figure 3.17: Boxplots of Relative L1 values for two to five clones at two tumour fraction levels in semi-realistic experiments. Datasets had a read depth of 1x and ten replicates. Superimposed strip plots assist in delineating model types.

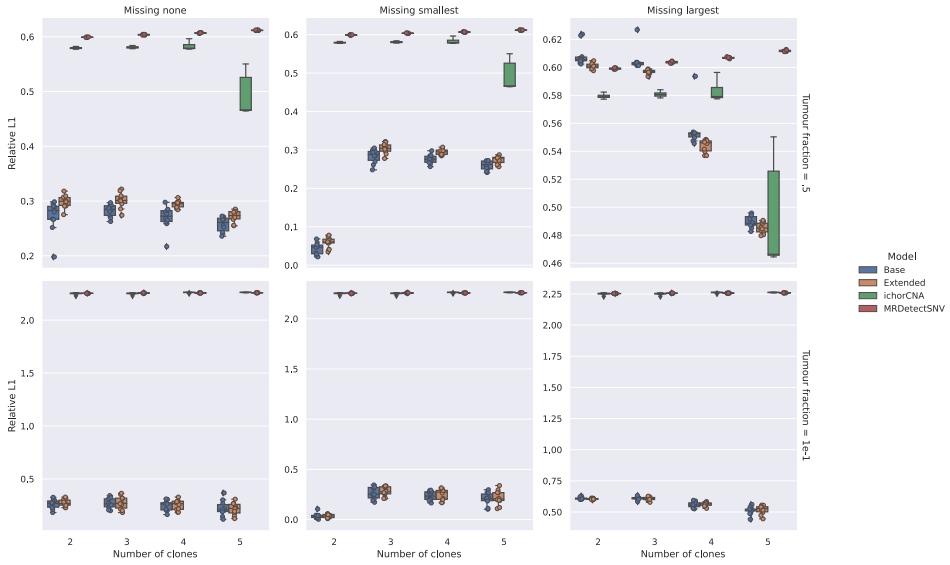


Figure 3.18: Boxplots of Relative L1 values for two to five clones with superimposed swarmplots comparing the effect of removing the smallest or largest clone in semi-realistic experiments. Table 3.10 documents the proportions of the removed clones. Datasets had a read depth of 1x.

Model	Tumour fraction	Number of clones	L1	Relative L1
Base	.5	2	0.156730	0.272266
		3	0.163208	0.282422
		4	0.154688	0.269335
		5	0.146814	0.257365
	1e-1	2	0.030042	0.261773
		3	0.031260	0.270887
		4	0.028189	0.247257
		5	0.025282	0.222929
Extended	.5	2	0.174267	0.298954
		3	0.175845	0.301263
		4	0.171004	0.294146
		5	0.156524	0.272298
	1e-1	2	0.031486	0.273173
		3	0.031857	0.274746
		4	0.028925	0.252938
		5	0.024392	0.216794
ichorCNA	.5	2	0.392530	0.579451
		3	0.393600	0.580649
		4	0.395117	0.582330
		5	0.317000	0.490334
	1e-1	2	0.850133	2.251404
		3	0.849707	2.250956
		4	0.859599	2.261340
		5	0.860352	2.262128
MRDetectSNV	.5	2	0.410336	0.599205
		3	0.414455	0.603720
		4	0.417405	0.606941
		5	0.421862	0.611788
	1e-1	2	0.852153	2.253555
		3	0.854658	2.256182
		4	0.855545	2.257111
		5	0.857088	2.258725

Table 3.9: Mean L1 and Relative L1 values across ten replicates for semi-realistic experiments on LiquidBayes' base and extended models, ichorCNA and MRDetectSNV for tw to five clones and two tumour fractions. Relative L1=log(L1 / tumour fraction + 1)

Number of Clones	Tumour fraction	Smallest	Largest
2	.5	0.1	0.4
	1e-1	0.02	0.08
3	.5	0.025	0.4
	1e-1	0.005	0.08
4	.5	0.025	0.35
	1e-1	0.005	0.07
5	.5	0.025	0.3
	1e-1	0.005	0.06

Table 3.10: Proportions of the smallest and largest clones that were removed in semi-realistic experiments at different numbers of clones and tumour fractions.

Chapter 4

Future Directions

Single Nucleotide Polymorphisms (SNPS) are genome positions with two distinct alleles that appear in a significant portion of the human population [32]. There are about 10 million SNPS in the human genome. SNPS comprise a major part of genetic variation among individuals [31] and explain differences in disease susceptibility. In future work, we intend to add SNPS into LiquidBayes. SNPS enable us to obtain allele-specific read counts, potentially increasing LiquidBayes' prediction power and sensitivity. This requires extracting SNPS from DLP+ results and blending an additional SNPS component in the model.

Further benchmarking of LiquidBayes can reveal additional insights into its strengths and weaknesses. We intend to perform experiments on cancer types other than Lymphoma to understand how LiquidBayes generalizes across cancers. In particular, we want to explore cancers with less aneuploidy, to investigate the utility of SNVS in our extended model. Due to time constraints, we were not able to run any experiments on real CTDNA data. This is an important task for future analyses.

Chapter 5

Discussion

Monitoring treatment efficacy can improve patient prognosis by detecting resistance and identifying MRD. Traditional tissue biopsies are invasive, thus prohibiting serial sampling. In contrast, liquid biopsies are non-invasive blood draws, making it a viable method for supervising post-treatment patient recovery. We present LiquidBayes, a BN which integrates clone-specific CN profiles from single-cell sequencing of the primary tumour in subsequent CTDNA analysis. Moreover, we propose a model extension where SNVS are included alongside CN profiles.

Post-treatment cancer patients exhibit very low tumour burdens, and often experience relapse later on. We showed that LiquidBayes was able to infer tumour burden at fractions as low as 1e-7 more accurately than the current SOTA (Section 3.2.1). Furthermore, LiquidBayes deconvolved individual clone proportions, providing insight into the clone-level mechanisms of resistance in patients. Specifically, by uncovering changes in clone populations using LiquidBayes, clinicians can identify resistant clones, informing treatment decisions and enhancing patient survival.

Read depth acts as a proxy for the signal-to-noise ratio, but higher read depths correspond to higher costs. In our experiments, LiquidBayes required a read depth of at least 1e-1x to outperform ichorCNA and MRDetectSNV (Section 3.2.2). Furthermore, higher read depths improved LiquidBayes' performance, whereas higher read depths did not affect ichorCNA or MRDetectSNV's performance.

Tumour heterogeneity brings about multiple clonal populations in the tumour. Moreover, the number of clonal populations across patients can differ considerably. LiquidBayes was robust to the number of clones, enabling it to generalize to cancers with diverse clone quantities. Cancer is also a dynamic process, giving rise to novel clonal populations over time. Hence, the clonal population structure in the initial tissue biopsy may not reflect the tumour at a future time point. Accordingly, we conducted experiments where we removed the smallest and largest clone from the CN profiles inputted to LiquidBayes. We found that LiquidBayes still outperformed other methods even when the smallest clone was removed (Section 3.2.4). The only instance where other methods gave better results than LiquidBayes was at a high tumour fraction (.5) and a low number of clones (< 4) (Section 3.2.4). However, post-treatment cancer patients do not exhibit high tumour burdens such as .5, making this case inapplicable for MRD settings.

In both the synthetic and semi-realistic experiments, we did not observe any significant improvement from the extended model over the base model. This could be an artifact of our SNV processing procedure or experimental design. Our SNV processing only left 100-200 SNVS, which may not be enough to contribute any additional power in our framework (Section 2.1.2). Regarding our experiments, the CN profiles of our sample manifested aneuploidy, suggesting CN to explain much of the genetic variation among clones.

In conclusion, LiquidBayes is a method which improves upon the current SOTA in liquid biopsy analysis. Our experiments showed that LiquidBayes outperformed SOTA methods at various tumour fractions, read depths and number of clones. Moreover, we showed that LiquidBayes was robust to missing clones in the initial tissue biopsy. LiquidBayes can be used in routine monitoring of cancer patients for detecting MRD.

Bibliography

[1] V. A. Adalsteinsson, G. Ha, S. S. Freeman, A. D. Choudhury, D. G. Stover, H. A. Parsons, G. Gydush, S. C. Reed, D. Rotem, J. Rhoades, D. Loginov, D. Livitz, D. Rosebrock, I. Leshchiner, J. Kim, C. Stewart, M. Rosenberg, J. M. Francis, C.-Z. Zhang, O. Cohen, C. Oh, H. Ding, P. Polak, M. Lloyd, S. Mahmud, K. Helvie, M. S. Merrill, R. A. Santiago, E. P. O'Connor, S. H. Jeong, R. Leeson, R. M. Barry, J. F. Kramkowski, Z. Zhang, L. Polacek, J. G. Lohr, M. Schleicher, E. Lipscomb, A. Saltzman, N. M. Oliver, L. Marini, A. G. Waks, L. C. Harshman, S. M. Tolaney, E. M. V. Allen, E. P. Winer, N. U. Lin, M. Nakabayashi, M.-E. Taplin, C. M. Johannessen, L. A. Garraway, T. R. Golub, J. S. Boehm, N. Wagle, G. Getz, J. C. Love, and M. Meyerson. Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. *Nature Communications*, 8(1), nov 2017. doi:10.1038/s41467-017-00965-y. → pages 1, 3, 21, 22

[2] C. Andrieu, N. de Freitas, A. Doucet, and M. I. Jordan. An introduction to mcmc for machine learning. *Machine Learning*, 50(1/2):5–43, 2003. doi:10.1023/a:1020281327116. → page 6

[3] P. L. Bedard, A. R. Hansen, M. J. Ratain, and L. L. Siu. Tumour heterogeneity in the clinic. *Nature*, 501(7467):355–364, sep 2013. doi:10.1038/nature12627. → page 2

[4] E. Bingham, J. P. Chen, M. Jankowiak, F. Obermeyer, N. Pradhan, T. Karaletsos, R. Singh, P. Szerlip, P. Horsfall, and N. D. Goodman. Pyro: Deep universal probabilistic programming, 2018.

[5] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 2008. ISBN 9780387310732. → page 6

[6] X. Chen, O. Schulz-Trieglaff, R. Shaw, B. Barnes, F. Schlesinger, M. Källberg, A. J. Cox, S. Kruglyak, and C. T. Saunders. Manta: rapid detection of structural variants and indels for germline and cancer

sequencing applications. *Bioinformatics*, 32(8):1220–1222, dec 2015. doi:10.1093/bioinformatics/btv710. → page 22

[7] R. B. Corcoran and B. A. Chabner. Application of cell-free DNA analysis to cancer treatment. *New England Journal of Medicine*, 379(18):1754–1765, nov 2018. doi:10.1056/nejmra1706174. → pages 1, 3

[8] S. Cristiano, A. Leal, J. Phallen, J. Fiksel, V. Adleff, D. C. Bruhm, S. Ø. Jensen, J. E. Medina, C. Hruban, J. R. White, D. N. Palsgrove, N. Niknafs, V. Anagnostou, P. Forde, J. Naidoo, K. Marrone, J. Brahmer, B. D. Woodward, H. Husain, K. L. van Rooijen, M.-B. W. Ørntoft, A. H. Madsen, C. J. H. van de Velde, M. Verheij, A. Cats, C. J. A. Punt, G. R. Vink, N. C. T. van Grieken, M. Koopman, R. J. A. Fijneman, J. S. Johansen, H. J. Nielsen, G. A. Meijer, C. L. Andersen, R. B. Scharpf, and V. E. Velculescu. Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature*, 570(7761):385–389, may 2019. doi:10.1038/s41586-019-1272-6. → page 21

[9] I. Dagogo-Jack and A. T. Shaw. Tumour heterogeneity and resistance to cancer therapies. *Nature Reviews Clinical Oncology*, 15(2):81–94, nov 2017. doi:10.1038/nrclinonc.2017.166. → page 2

[10] P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, and R. D. and. The variant call format and VCFtools. *Bioinformatics*, 27(15): 2156–2158, jun 2011. doi:10.1093/bioinformatics/btr330. → page 22

[11] P. Danecek, J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, T. Keane, S. A. McCarthy, R. M. Davies, and H. Li. Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2), jan 2021. doi:10.1093/gigascience/giab008. → page 19

[12] V. der Auwera GA and O. BD. Genomics in the cloud: Using docker, gatk, and wdl in terra (1st edition). *O'Reilly Media*, 2020. → page 19

[13] J. Donaldson and B. H. Park. Circulating tumor DNA: Measurement and clinical utility. *Annual Review of Medicine*, 69(1):223–234, jan 2018. doi:10.1146/annurev-med-041316-085721. → page 1

[14] P. Eirew, A. Steif, J. Khattra, G. Ha, D. Yap, H. Farahani, K. Gelmon, S. Chia, C. Mar, A. Wan, E. Laks, J. Biele, K. Shumansky, J. Rosner, A. McPherson, C. Nielsen, A. J. L. Roth, C. Lefebvre, A. Bashashati, C. de Souza, C. Siu, R. Aniba, J. Brimhall, A. Oloumi, T. Osako, A. Bruna,

J. L. Sandoval, T. Algara, W. Greenwood, K. Leung, H. Cheng, H. Xue, Y. Wang, D. Lin, A. J. Mungall, R. Moore, Y. Zhao, J. Lorette, L. Nguyen, D. Huntsman, C. J. Eaves, C. Hansen, M. A. Marra, C. Caldas, S. P. Shah, and S. Aparicio. Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. *Nature*, 518(7539):422–426, nov 2014. doi:10.1038/nature13952. → page 3

[15] L. Feuk, A. R. Carson, and S. W. Scherer. Structural variation in the human genome. *Nature Reviews Genetics*, 7(2):85–97, feb 2006. doi:10.1038/nrg1767. → page 4

[16] A. Fischer, I. Vázquez-García, C. J. Illingworth, and V. Mustonen. High-definition reconstruction of clonal composition in cancer. *Cell Reports*, 7(5):1740–1752, jun 2014. doi:10.1016/j.celrep.2014.04.055. → page 4

[17] A. E. Gelfand and A. F. M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410): 398–409, jun 1990. doi:10.1080/01621459.1990.10476213. → page 7

[18] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741, nov 1984. doi:10.1109/tpami.1984.4767596. → page 7

[19] M. Gerstung, C. Beisel, M. Rechsteiner, P. Wild, P. Schraml, H. Moch, and N. Beerenwinkel. Reliable detection of subclonal single-nucleotide variants in tumour cell populations. *Nature Communications*, 3(1), jan 2012. doi:10.1038/ncomms1814. → page 3

[20] N. D. Goodman. The principles and practice of probabilistic programming. *ACM SIGPLAN Notices*, 48(1):399–402, jan 2013. doi:10.1145/2480359.2429117.

[21] G. Ha, A. Roth, J. Khattra, J. Ho, D. Yap, L. M. Prentice, N. Melnyk, A. McPherson, A. Bashashati, E. Laks, J. Biele, J. Ding, A. Le, J. Rosner, K. Shumansky, M. A. Marra, C. B. Gilks, D. G. Huntsman, J. N. McAlpine, S. Aparicio, and S. P. Shah. TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Research*, 24(11):1881–1893, jul 2014. doi:10.1101/gr.180281.114. → pages 2, 4

[22] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, apr 1970. doi:10.1093/biomet/57.1.97. → page 7

[23] M. J. Higgins, D. Jelovac, E. Barnathan, B. Blair, S. Slater, P. Powers, J. Zorzi, S. C. Jeter, G. R. Oliver, J. Fetting, L. Emens, C. Riley, V. Stearns, F. Diehl, P. Angenendt, P. Huang, L. Cope, P. Argani, K. M. Murphy, K. E. Bachman, J. Greshock, A. C. Wolff, and B. H. Park. Detection of tumor pik3ca status in metastatic breast cancer using peripheral blood. *Clinical Cancer Research*, 18(12):3462–3469, jun 2012. doi:10.1158/1078-0432.ccr-11-2696. → page 1

[24] M. D. Hoffman and A. Gelman. The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. 2011. doi:10.48550/ARXIV.1111.4246. → page 14

[25] R. J. Hyndman. Computing and graphing highest density regions. *The American Statistician*, 50(2):120, may 1996. doi:10.2307/2684423. → page 7

[26] M. Ilié and P. Hofman. Pros: Can tissue biopsy be replaced by liquid biopsy? *Translational Lung Cancer Research*, 5(4):420–423, aug 2016. doi:10.21037/tlcr.2016.08.06. → page 2

[27] S. Kang, Q. Li, Q. Chen, Y. Zhou, S. Park, G. Lee, B. Grimes, K. Krysan, M. Yu, W. Wang, F. Alber, F. Sun, S. M. Dubinett, W. Li, and X. J. Zhou. CancerLocator: non-invasive cancer diagnosis and tissue-of-origin prediction using methylation profiles of cell-free DNA. *Genome Biology*, 18 (1), mar 2017. doi:10.1186/s13059-017-1191-5. → page 3

[28] S. Kim, K. Scheffler, A. L. Halpern, M. A. Bekritsky, E. Noh, M. Källberg, X. Chen, Y. Kim, D. Beyter, P. Krusche, and C. T. Saunders. Strelka2: fast and accurate calling of germline and somatic variants. *Nature Methods*, 15 (8):591–594, jul 2018. doi:10.1038/s41592-018-0051-x. → page 22

[29] D. Koller. *Probabilistic graphical models*. MIT Press, 2010. ISBN 9780262013192. → page 6

[30] C. Krapu and M. Borsuk. Probabilistic programming: A review for environmental modellers. *Environmental Modelling & Software*, 114:40–48, apr 2019. doi:10.1016/j.envsoft.2019.01.014. → page 6

[31] L. Kruglyak and D. A. Nickerson. Variation is the spice of life. *Nature Genetics*, 27(3):234–236, mar 2001. doi:10.1038/85776. → page 51

[32] T. LaFramboise. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic Acids Research*, 37(13):4181–4193, jul 2009. doi:10.1093/nar/gkp552. → page 51

[33] D. Lai. Hmm copy utils. *GitHub repository*, 2011. → pages 10, 21, 22

[34] D. Lai, G. Ha, and S. Shah. Hmmcopy: Copy number prediction with correction for gc and mappability bias for hts data. 2021. R package version 1.36.0. → pages 4, 10, 19

[35] E. Lakatos, H. Hockings, M. Mossner, W. Huang, M. Lockley, and T. A. Graham. LiquidCNA: Tracking subclonal evolution from longitudinal liquid biopsies using somatic copy number alterations. *iScience*, 24(8):102889, aug 2021. doi:10.1016/j.isci.2021.102889. → page 3

[36] E. Laks, A. McPherson, H. Zahn, D. Lai, A. Steif, J. Brimhall, J. Biele, B. Wang, T. Masud, J. Ting, D. Grewal, C. Nielsen, S. Leung, V. Bojilova, M. Smith, O. Golovko, S. Poon, P. Eirew, F. Kabeer, T. R. de Algara, S. R. Lee, M. J. Taghiyar, C. Huebner, J. Ngo, T. Chan, S. Vatrt-Watts, P. Walters, N. Abrar, S. Chan, M. Wiens, L. Martin, R. W. Scott, T. M. Underhill, E. Chavez, C. Steidl, D. D. Costa, Y. Ma, R. J. Coope, R. Corbett, S. Pleasance, R. Moore, A. J. Mungall, C. Mar, F. Cafferty, K. Gelmon, S. Chia, M. A. Marra, C. Hansen, S. P. Shah, S. Aparicio, G. J. Hannon, G. Battistoni, D. Bressan, I. Cannell, H. Casbolt, C. Jauset, T. Kovačević, C. Mulvey, F. Nugent, M. P. Ribes, I. Pearsall, F. Qosaj, K. Sawicka, S. Wild, E. Williams, S. Aparicio, E. Laks, Y. Li, C. O'Flanagan, A. Smith, T. Ruiz, S. Balasubramanian, M. Lee, B. Bodenmiller, M. Burger, L. Kuett, S. Tietscher, J. Windager, E. Boyden, S. Alon, Y. Cui, A. Emenari, D. Goodwin, E. Karagiannis, A. Sinha, A. T. Wassie, C. Caldas, A. Bruna, M. Callari, W. Greenwood, G. Lerda, Y. Lubling, A. Marti, O. Rueda, A. Shea, O. Harris, R. Becker, F. Grimaldi, S. Harris, S. Vogl, J. A. Joyce, J. Hausser, S. Watson, S. Shah, A. McPherson, I. Vázquez-García, S. Tavaré, K. Dinh, E. Fisher, R. Kunes, N. A. Walton, M. A. Sa'd, N. Chornay, A. Dariush, E. G. Solares, C. Gonzalez-Fernandez, A. K. Yoldas, N. Millar, X. Zhuang, J. Fan, H. Lee, L. S. Duran, C. Xia, and P. Zheng. Clonal decomposition and DNA replication states defined by scaled single-cell genome sequencing. *Cell*, 179(5):1207–1221.e22, nov 2019. doi:10.1016/j.cell.2019.10.026. → pages 4, 5, 19

[37] J. Li, L. Wei, X. Zhang, W. Zhang, H. Wang, B. Zhong, Z. Xie, H. Lv, and X. Wang. DISMIR: Deep learning-based noninvasive cancer detection by integrating DNA sequence and methylation information of individual

cell-free DNA reads. *Briefings in Bioinformatics*, 22(6), jul 2021. doi:10.1093/bib/bbab250. → page 3

[38] L. McInnes, J. Healy, and S. Astels. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2(11):205, mar 2017. doi:10.21105/joss.00205.

[39] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2018. → page 4

[40] A. W. McPherson, A. Roth, G. Ha, C. Chauve, A. Steif, C. P. E. de Souza, P. Eirew, A. Bouchard-Côté, S. Aparicio, S. C. Sahinalp, and S. P. Shah. ReMixT: clone-specific genomic structure estimation in cancer. *Genome Biology*, 18(1), jul 2017. doi:10.1186/s13059-017-1267-2. → pages 2, 4

[41] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, jun 1953. doi:10.1063/1.1699114. → page 7

[42] F. Mölder, K. P. Jablonski, B. Letcher, M. B. Hall, C. H. Tomkins-Tinch, V. Sochat, J. Forster, S. Lee, S. O. Twardziok, A. Kanitz, A. Wilm, M. Holtgrewe, S. Rahmann, S. Nahnsen, and J. Köster. Sustainable data analysis with snakemake. *F1000Research*, 10:33, apr 2021. doi:10.12688/f1000research.29032.2. → page 21

[43] L. Müllauer. Next generation sequencing: clinical applications in solid tumours. *memo - Magazine of European Medical Oncology*, 10(4):244–247, nov 2017. doi:10.1007/s12254-017-0361-1. → page 3

[44] T. Nawy. Single-cell sequencing. *Nature Methods*, 11(1):18–18, dec 2013. doi:10.1038/nmeth.2771. → page 3

[45] S. Nik-Zainal, L. B. Alexandrov, D. C. Wedge, P. V. Loo, C. D. Greenman, K. Raine, D. Jones, J. Hinton, J. Marshall, L. A. Stebbings, A. Menzies, S. Martin, K. Leung, L. Chen, C. Leroy, M. Ramakrishna, R. Rance, K. W. Lau, L. J. Mudie, I. Varela, D. J. McBride, G. R. Bignell, S. L. Cooke, A. Shlien, J. Gamble, I. Whitmore, M. Maddison, P. S. Tarpey, H. R. Davies, E. Papaemmanuil, P. J. Stephens, S. McLaren, A. P. Butler, J. W. Teague, G. Jönsson, J. E. Garber, D. Silver, P. Miron, A. Fatima, S. Boyault, A. Langerød, A. Tutt, J. W. Martens, S. A. Aparicio, Å. Borg, A. V. Salomon, G. Thomas, A.-L. Børresen-Dale, A. L. Richardson, M. S. Neuberger, P. A. Futreal, P. J. Campbell, and M. R. Stratton. Mutational

processes molding the genomes of 21 breast cancers. *Cell*, 149(5):979–993, may 2012. doi:10.1016/j.cell.2012.04.024. → page 1

[46] P. C. Nowell. The clonal evolution of tumor cell populations. *Science*, 194 (4260):23–28, oct 1976. doi:10.1126/science.959840. → pages 1, 2

[47] L. Oesper, G. Satas, and B. J. Raphael. Quantifying tumor heterogeneity in whole-genome and whole-exome sequencing data. *Bioinformatics*, 30(24): 3532–3540, oct 2014. doi:10.1093/bioinformatics/btu651. → pages 2, 4

[48] M. P and M. P. Les acides nucléiques du plasma sanguin chez l'homme. *C R Seances Soc Biol Fil*, (142):241–243, 1948. → page 2

[49] D. Phan, N. Pradhan, and M. Jankowiak. Composable effects for flexible and accelerated probabilistic programming in numpyro. 2019. doi:10.48550/ARXIV.1912.11554. → page 14

[50] P. Ramos and M. Bentires-Alj. Mechanism-based cancer therapy: resistance to therapy, therapy for resistance. *Oncogene*, 34(28):3617–3626, sep 2014. doi:10.1038/onc.2014.314. → pages 1, 2

[51] A. Roth, J. Khattra, D. Yap, A. Wan, E. Laks, J. Biele, G. Ha, S. Aparicio, A. Bouchard-Côté, and S. P. Shah. PyClone: statistical inference of clonal population structure in cancer. *Nature Methods*, 11(4):396–398, mar 2014. doi:10.1038/nmeth.2883. → page 2

[52] S. Salehi, F. Dorri, K. Chern, F. Kabeer, N. Rusk, T. Funnell, M. J. Williams, D. Lai, M. Andronescu, K. R. Campbell, A. McPherson, S. Aparicio, A. Roth, S. P. Shah, and A. Bouchard-Côté. Cancer phylogenetic tree inference at scale from 1000s of single cell genomes. may 2020. doi:10.1101/2020.05.06.058180. → page 19

[53] E. Shapiro, T. Biezuner, and S. Linnarsson. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature Reviews Genetics*, 14(9):618–630, jul 2013. doi:10.1038/nrg3542. → page 3

[54] R. Shen and V. E. Seshan. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Research*, 44(16):e131–e131, jun 2016. doi:10.1093/nar/gkw520. → pages 2, 4

[55] G. Siravegna, S. Marsoni, S. Siena, and A. Bardelli. Integrating liquid biopsies into the management of cancer. *Nature Reviews Clinical Oncology*, 14(9):531–548, mar 2017. doi:10.1038/nrclinonc.2017.14. → page 2

[56] M. W. Snyder, M. Kircher, A. J. Hill, R. M. Daza, and J. Shendure. Cell-free DNA comprises an in vivo nucleosome footprint that informs its tissues-of-origin. *Cell*, 164(1-2):57–68, jan 2016. doi:10.1016/j.cell.2015.11.050. → page 21

[57] C. D. Steele, A. Abbasi, S. M. A. Islam, A. L. Bowes, A. Khandekar, K. Haase, S. Hames-Fathi, D. Ajayi, A. Verfaillie, P. Dhami, A. McLatchie, M. Lechner, N. Light, A. Shlien, D. Malkin, A. Feber, P. Proszek, T. Lesluyes, F. Mertens, A. M. Flanagan, M. Tarabichi, P. V. Loo, L. B. Alexandrov, and N. Pillay. Signatures of copy number alterations in human cancer. *Nature*, 606(7916):984–991, jun 2022. doi:10.1038/s41586-022-04738-6. → page 4

[58] D. Sylvie-Louise Avon and H. Klieb. Oral soft-tissue biopsy: an overview. *J Can Dent Assoc*, 78:c75, 2012. → page 2

[59] D. Tamborero, C. Rubio-Perez, J. Deu-Pons, M. P. Schroeder, A. Vivancos, A. Rovira, I. Tusquets, J. Albanell, J. Rodon, J. Tabernero, C. de Torres, R. Dienstmann, A. Gonzalez-Perez, and N. Lopez-Bigas. Cancer genome interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Medicine*, 10(1), mar 2018. doi:10.1186/s13073-018-0531-8. → page 5

[60] C. Tomasetti, L. Marchionni, M. A. Nowak, G. Parmigiani, and B. Vogelstein. Only three driver gene mutations are required for the development of lung and colorectal cancers. *Proceedings of the National Academy of Sciences*, 112(1):118–123, dec 2014. doi:10.1073/pnas.1421839112. → page 5

[61] J.-W. van de Meent, B. Paige, H. Yang, and F. Wood. An introduction to probabilistic programming, 2018.

[62] S. Volik, M. Alcaide, R. D. Morin, and C. Collins. Cell-free DNA (cfDNA): Clinical significance and utility in cancer shaped by emerging technologies. *Molecular Cancer Research*, 14(10):898–908, oct 2016. doi:10.1158/1541-7786.mcr-16-0044. → page 2

[63] J. C. M. Wan, C. Massie, J. Garcia-Corbacho, F. Mouliere, J. D. Brenton, C. Caldas, S. Pacey, R. Baird, and N. Rosenfeld. Liquid biopsies come of age: towards implementation of circulating tumour DNA. *Nature Reviews Cancer*, 17(4):223–238, feb 2017. doi:10.1038/nrc.2017.7. → page 1

[64] R. Xi, A. G. Hadjipanayis, L. J. Luquette, T.-M. Kim, E. Lee, J. Zhang, M. D. Johnson, D. M. Muzny, D. A. Wheeler, R. A. Gibbs, R. Kucherlapati, and P. J. Park. Copy number variation detection in whole-genome sequencing data using the bayesian information criterion. *Proceedings of the National Academy of Sciences*, 108(46), nov 2011. doi:10.1073/pnas.1110574108. → page 4

[65] C. Xu. A review of somatic single nucleotide variant calling algorithms for next-generation sequencing data. *Computational and Structural Biotechnology Journal*, 16:15–24, 2018. doi:10.1016/j.csbj.2018.01.003. → page 5

[66] H. Zahn, A. Steif, E. Laks, P. Eirew, M. VanInsberghe, S. P. Shah, S. Aparicio, and C. L. Hansen. Scalable whole-genome single-cell library preparation without preamplification. *Nature Methods*, 14(2):167–173, jan 2017. doi:10.1038/nmeth.4140. → page 3

[67] C. Zong, S. Lu, A. R. Chapman, and X. S. Xie. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science*, 338(6114):1622–1626, dec 2012. doi:10.1126/science.1229164. → page 3

[68] H. Zou, L.-X. Wu, L. Tan, F.-F. Shang, and H.-H. Zhou. Significance of single-nucleotide variants in long intergenic non-protein coding RNAs. *Frontiers in Cell and Developmental Biology*, 8, may 2020. doi:10.3389/fcell.2020.00347. → page 5

[69] A. Zviran, R. C. Schulman, M. Shah, S. T. K. Hill, S. Deochand, C. C. Khamnei, D. Maloney, K. Patel, W. Liao, A. J. Widman, P. Wong, M. K. Callahan, G. Ha, S. Reed, D. Rotem, D. Frederick, T. Sharova, B. Miao, T. Kim, G. Gydush, J. Rhoades, K. Y. Huang, N. D. Omans, P. O. Bolan, A. H. Lipsky, C. Ang, M. Malbari, C. F. Spinelli, S. Kazancioglu, A. M. Runnels, S. Fennessey, C. Stolte, F. Gaiti, G. G. Inghirami, V. Adalsteinsson, B. Houck-Loomis, J. Ishii, J. D. Wolchok, G. Boland, N. Robine, N. K. Altorki, and D. A. Landau. Genome-wide cell-free DNA mutational integration enables ultra-sensitive cancer monitoring. *Nature Medicine*, 26 (7):1114–1124, jun 2020. doi:10.1038/s41591-020-0915-3. → pages 3, 21

Appendix A

Supporting Materials

A.1 Figures & Tables

Number of Clones	ρ
2	$(0.1, 0.9)$
3	$(0.1, 0.7, 0.2)$
4	$(0.05, 0.6, 0.2, 0.15)$
5	$(0.05, 0.4, 0.05, 0.25, 0.25)$
6	$(0.05, 0.4, 0.05, 0.15, 0.25, 0.15)$

Table A.1: Unnormalized ρ values at different numbers of clones. ρ was determined by the number of clones.

Tumour fraction	Clone	L1	Relative L1
.5	A (0.35)	0.000698	0.001394
	B (0.1)	0.000274	0.000548
	C (0.025)	0.000231	0.000463
	D (0.025)	0.000257	0.000514
.3	A (0.21)	0.000266	0.000887
	B (0.06)	0.000343	0.001143
	C (0.015)	0.000301	0.001003
	D (0.015)	0.000266	0.000886
1e-1	A (7e-2)	0.000316	0.003155
	B (2e-2)	0.000329	0.003277
	C (5e-3)	0.000198	0.001979
	D (5e-3)	0.000200	0.001994
1e-3	A (7e-4)	0.000729	0.547301
	B (2e-4)	0.000038	0.036935
	C (5e-5)	0.000124	0.116377
	D (5e-5)	0.000035	0.034057
1e-5	A (7e-6)	0.000033	1.297947
	B (2e-6)	0.000049	1.541569
	C (5e-7)	0.000038	1.338769
	D (5e-7)	0.000040	1.403804
1e-7	A (7e-8)	0.000039	5.264066
	B (2e-8)	0.000050	5.566370
	C (5e-9)	0.000043	5.382725
	D (5e-9)	0.000042	5.400277

Table A.2: Mean L1 and Relative L1 clone fraction estimates across ten replicates for synthetic experiments on the extended model at six tumour fraction levels. Ground truth proportion for clones are recorded in parenthesis beside the label in the clone column. Relative L1=log(L1 / tumour fraction + 1).

Tumour fraction	Read depth	L1	Relative L1
	100	0.000377	0.003756
	10	0.000411	0.004101
	1	0.000649	0.006447
1e-1	.5	0.001027	0.010205
	1e-1	0.002886	0.028274
	1e-2	0.007341	0.069998
	1e-3	0.016045	0.146775
	100	0.000038	0.003782
	10	0.000229	0.022587
	1	0.000726	0.069420
1e-2	.5	0.001092	0.100823
	1e-1	0.001923	0.168987
	1e-2	0.002499	0.208565
	1e-3	0.012454	0.766878
	100	0.000043	0.008541
	10	0.000101	0.019796
	1	0.000547	0.101063
5e-3	.5	0.000800	0.143213
	1e-1	0.001884	0.302335
	1e-2	0.003844	0.528497
	1e-3	0.012627	1.224582
	100	0.000902	0.641871
	10	0.000259	0.220185
	1	0.000882	0.631397
1e-3	.5	0.000810	0.591815
	1e-1	0.001738	0.887751
	1e-2	0.000746	0.554169
	1e-3	0.000796	0.584141

Table A.3: Mean L1 and Relative L1 across ten replicates for internal experiments on the extended model at seven read depths and 3 tumour fraction levels. Relative L1=log(L1 / tumour fraction + 1)

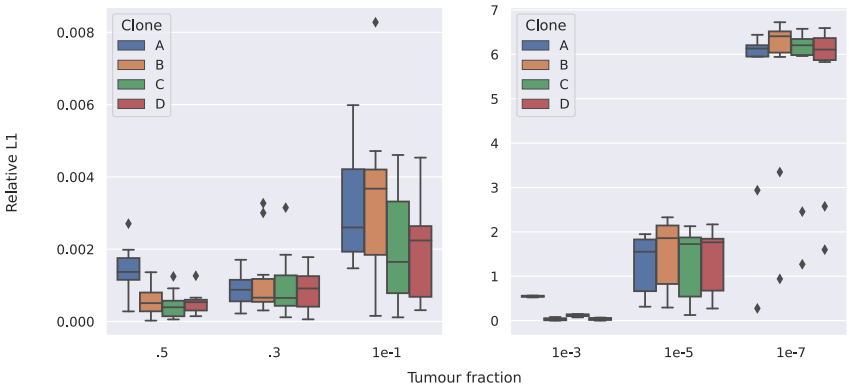
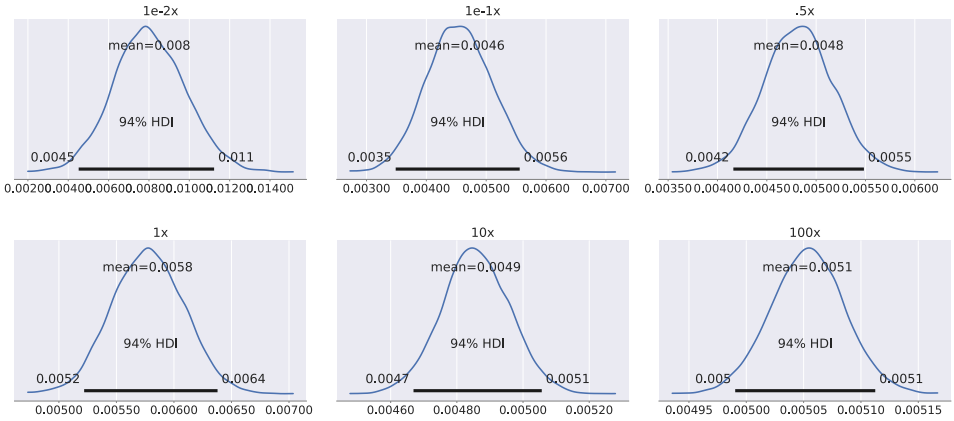


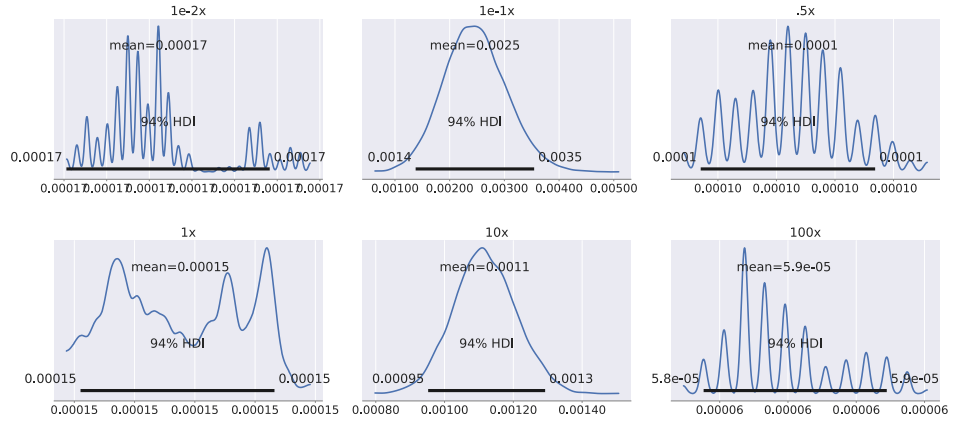
Figure A.1: Boxplots of Relative L1 values for tumour and clone fraction estimates for six tumour fraction levels using the extended model. Simulated datasets had three clones and ten replicates were generated.

Tumour fraction	Number of clones	L1	Relative L1
.5	2	0.003706	0.007363
	3	0.005544	0.011018
	4	0.000884	0.001766
	5	0.000731	0.001461
	6	0.000716	0.001431
1e-1	2	0.002185	0.021505
	3	0.000649	0.006447
	4	0.000348	0.003469
	5	0.000413	0.004114
	6	0.000405	0.004036

Table A.4: Mean L1 and Relative L1 across ten replicates for internal experiments on the extended model at five different numbers of clones and two tumour fraction levels. Relative L1=log(L1 / tumour fraction + 1)

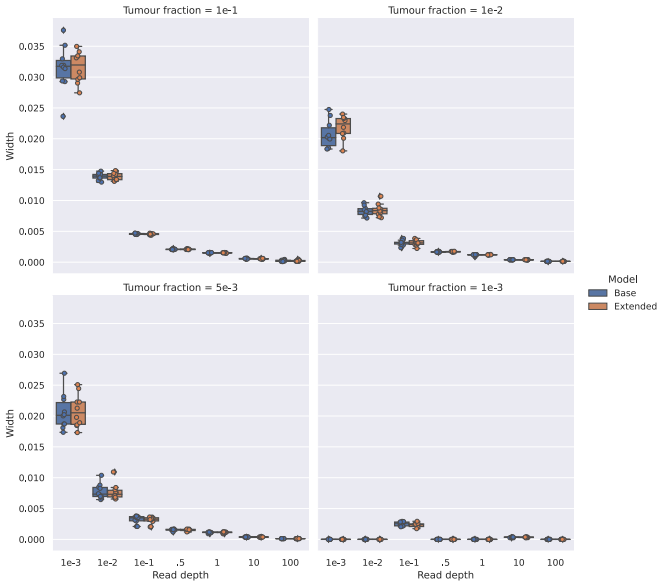


(a) Extended model posterior plots at Tumour fraction = 5e-3.

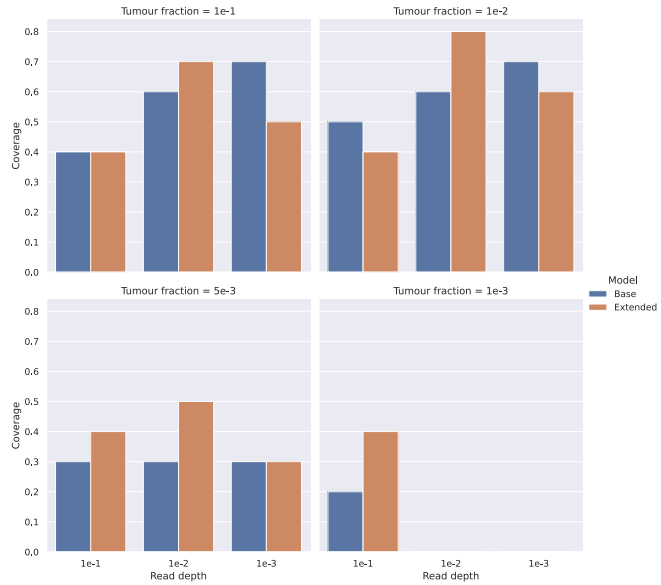


(b) Extended model posterior plots at Tumour fraction = 1e-3.

Figure A.2: Posterior plots from the extended model at six read depths for tumour fractions 5e-3 and 1e-3. Bold black lines indicate the HDI.

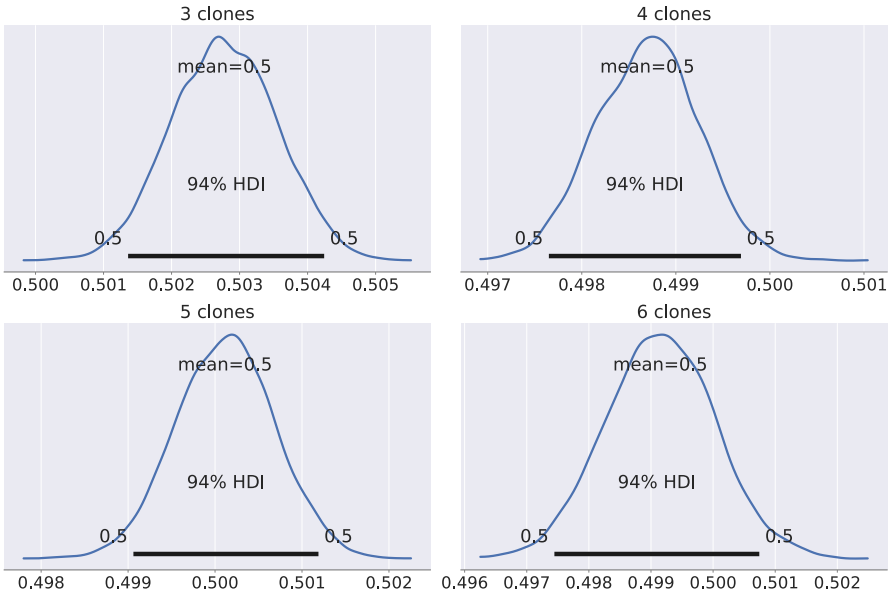


(a) Distribution of 94% HDI widths across ten replicates for the base and extended models at seven read depths and four tumour fractions. A strip plot is superimposed for easily identifying model types.

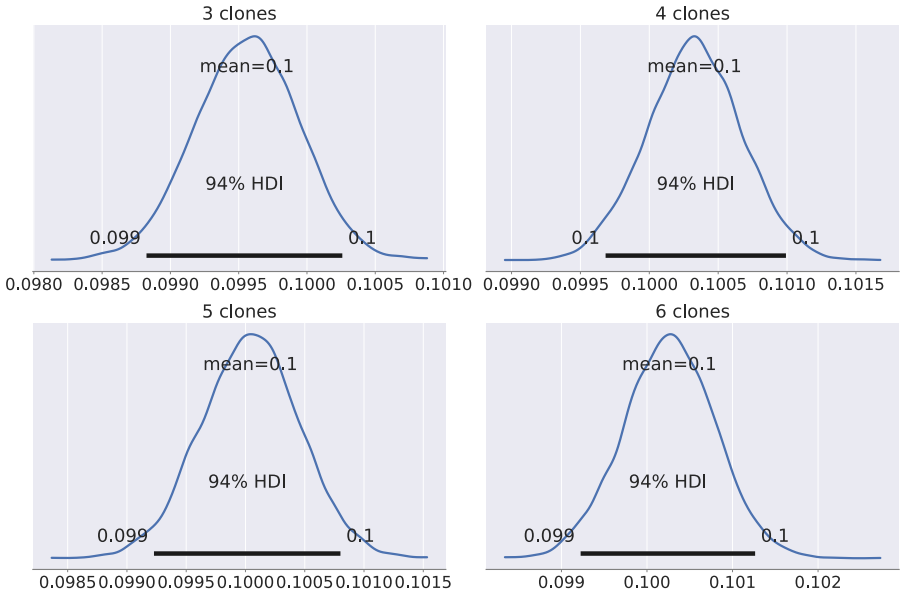


(b) Bar plot depicting the Bayesian coverage across ten replicates at three read depths and four tumour fractions.

Figure A.3: Posterior distribution statistics for synthetic experiments on read depth. **(a)** HDI widths. **(b)** Bayesian coverage.

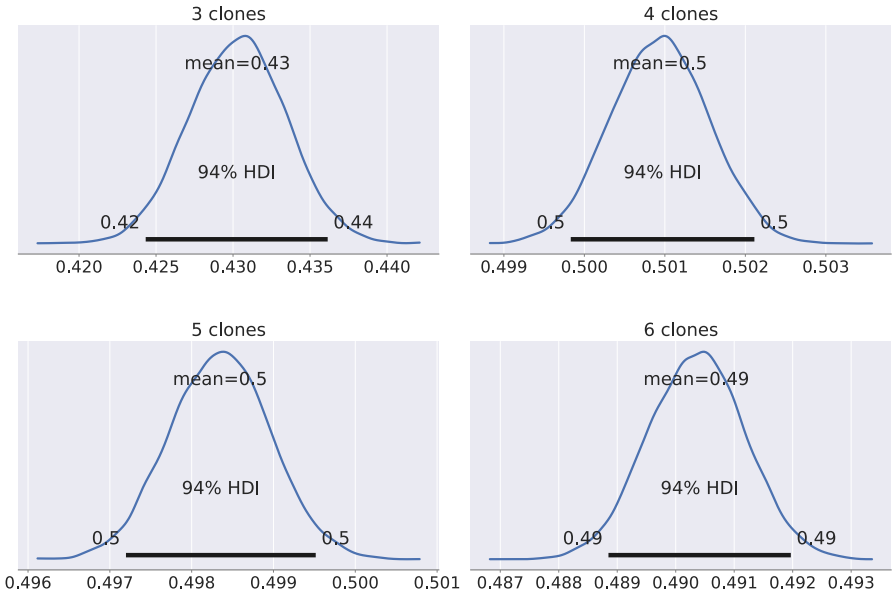


(a) Tumour fraction = .5.

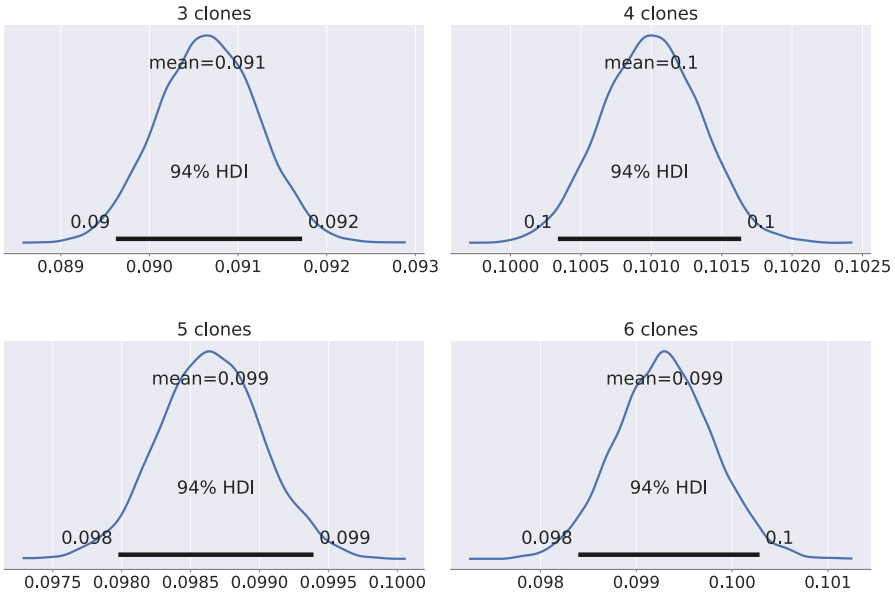


(b) Tumour fraction = 1e-1.

Figure A.4: Posterior plots from synthetic experiments using the extended model for three to six clones at tumour fractions .5 and 1e-1. Bold black lines indicate the HDI.

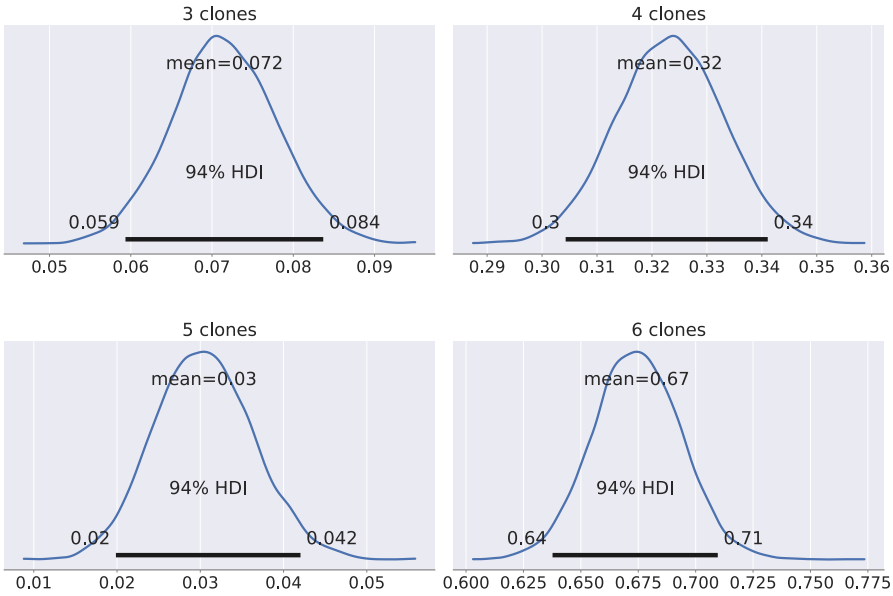


(a) Tumour fraction = .5.

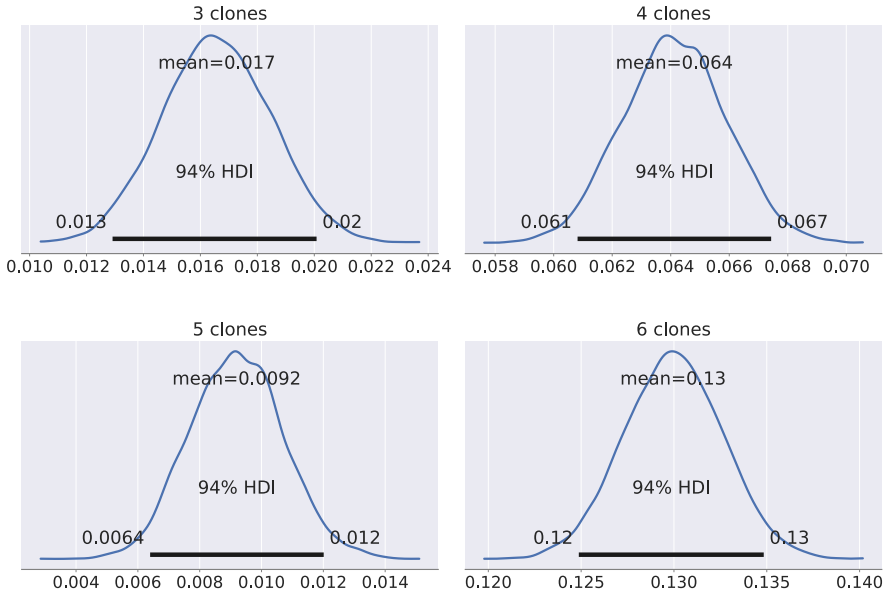


(b) Tumour fraction = 1e-1.

Figure A.5: Posterior plots from the base model for three to six clones at tumour fractions .5 and 1e-1, with the smallest clone removed. Bold black lines indicate the HDI.

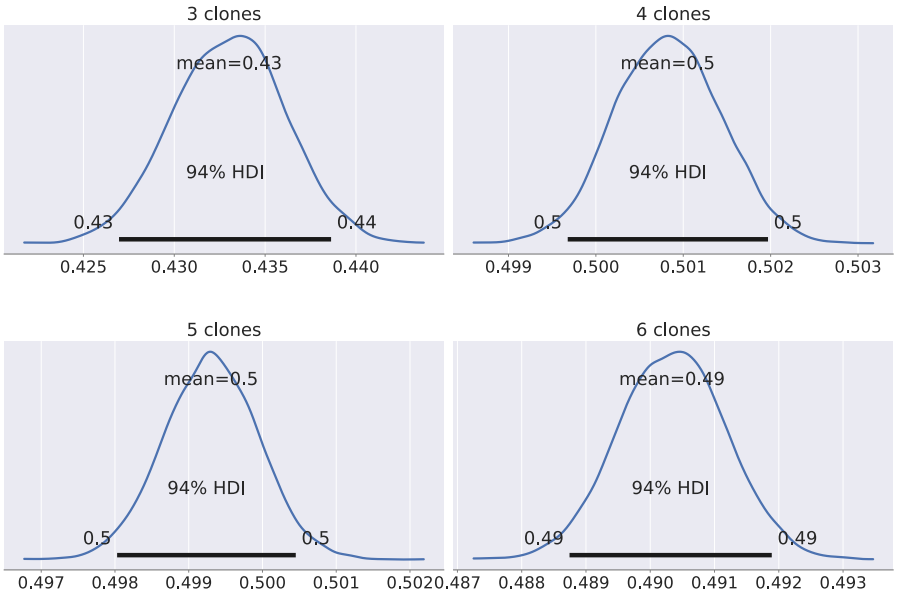


(a) Tumour fraction = .5.

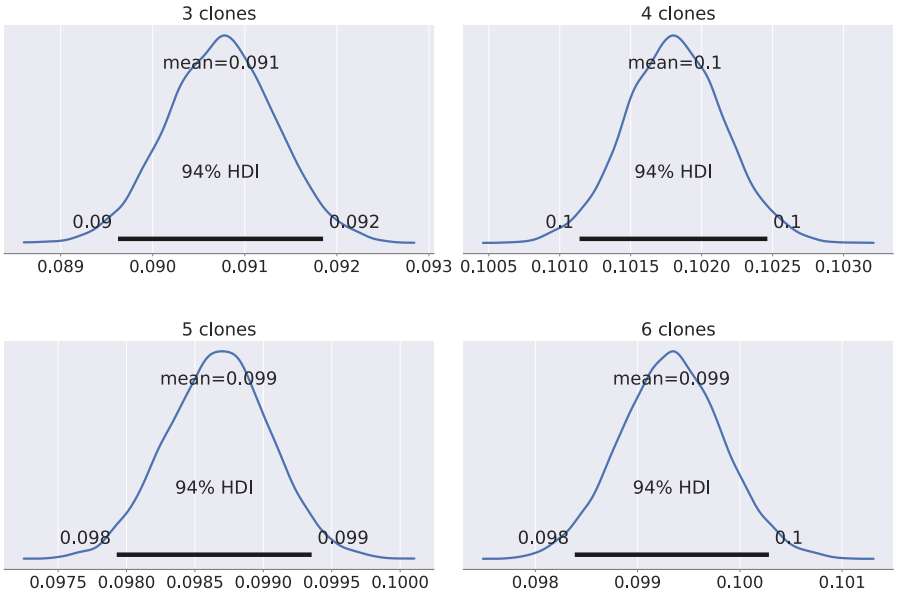


(b) Tumour fraction = 1e-1.

Figure A.6: Posterior plots from the base model for three to six clones at tumour fractions .5 and 1e-1, with the largest clone removed. Bold black lines indicate the HDI.

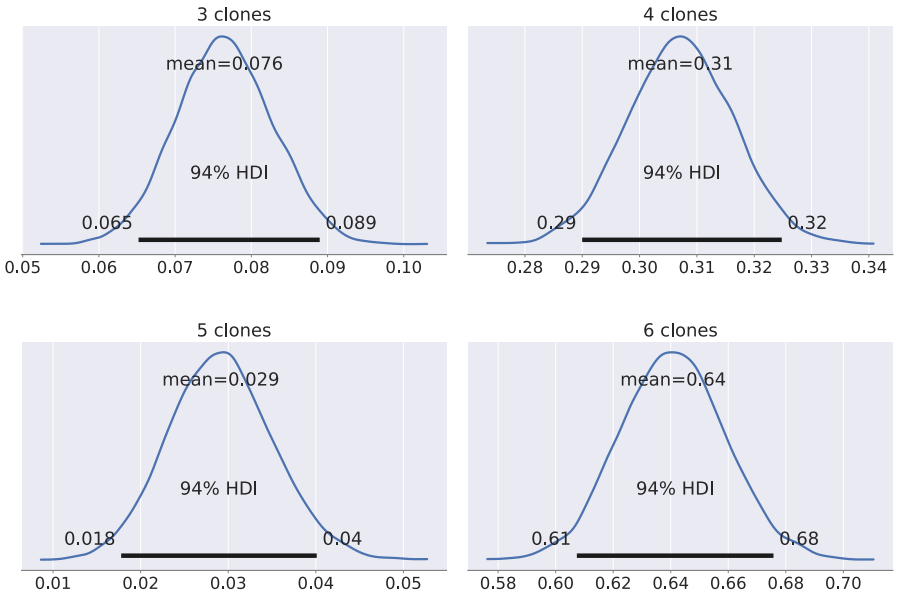


(a) Tumour fraction = .5.

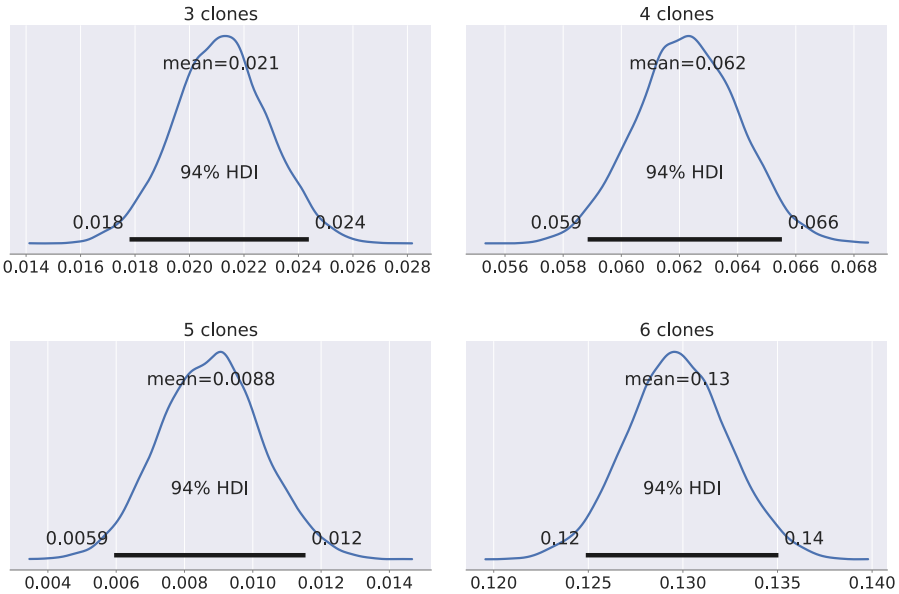


(b) Tumour fraction = 1e-1.

Figure A.7: Posterior plots from the extended model for two to six clones at tf=.5 and 1e-1, with the smallest clone removed. Bold black lines indicate the HDI



(a) Tumour fraction = .5.



(b) Tumour fraction = 1e-1.

Figure A.8: Posterior plots from the extended model for two to six clones at tf=1e-1, with the largest clone removed. Bold black lines indicate the HDI

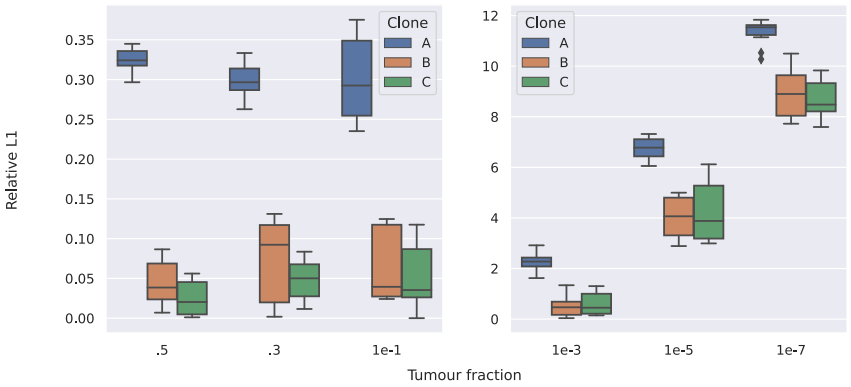
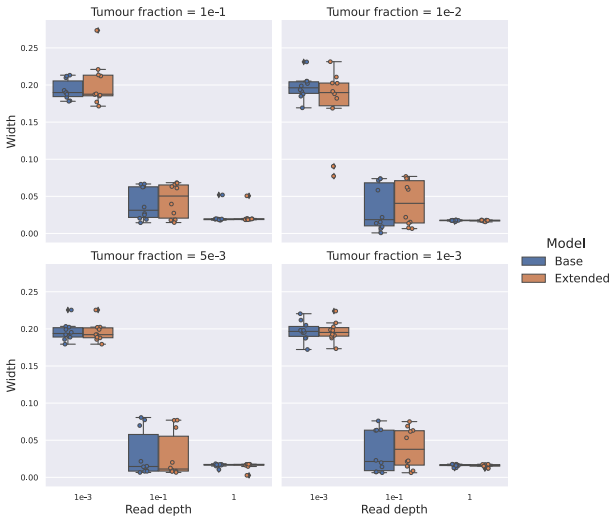


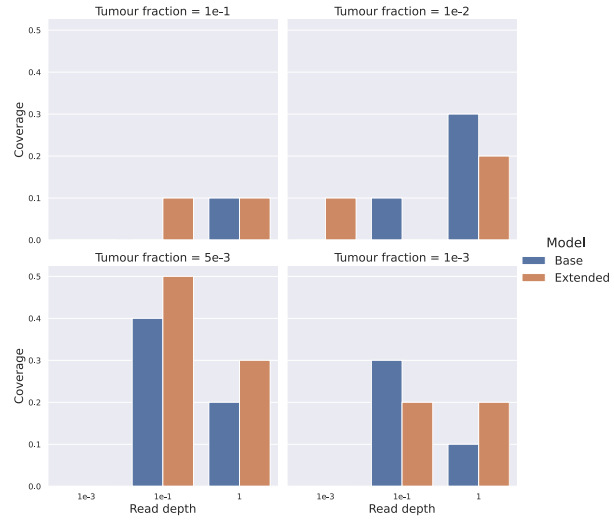
Figure A.9: Boxplots of Relative L1 values for tumour and clone fraction estimates on semi-realistic datasets for six tumour fraction levels using the extended model. Datasets had three clones A,B and C at proportions .8, .15 and .15, respectively and a read depth of 1x.

clone	tumour fraction	true proportion	L1	log(L1 / tumour fraction + 1)
A	.5	0.4	0.191953	0.324806
	.3	0.24	0.104157	0.297788
	1e-1	0.08	0.034942	0.298352
	1e-3	0.0008	0.009141	2.248056
	1e-5	8.e-06	0.009248	6.743687
	1e-7	8e-08	0.009250	11.332886
B	.5	0.075	0.022910	0.044471
	.3	0.045	0.023360	0.073753
	1e-1	0.015	0.006937	0.066098
	1e-3	0.00015	0.000762	0.487754
	1e-5	1.5e-06	0.000718	4.015869
	1e-7	1.5e-08	0.001186	8.918982
C	.5	0.025	0.012587	0.024641
	.3	0.015	0.015043	0.048627
	1e-1	0.005	0.005494	0.052807
	1e-3	5e-05	0.001011	0.605934
	1e-5	5.e-07	0.001294	4.242566
	1e-7	5e-09	0.000778	8.706282

Table A.5: Mean L1 and Relative L1 clone fraction estimates across ten replicates for synthetic experiments on the extended model at six tumour fraction levels. Relative L1=log(L1 / tumour fraction + 1).

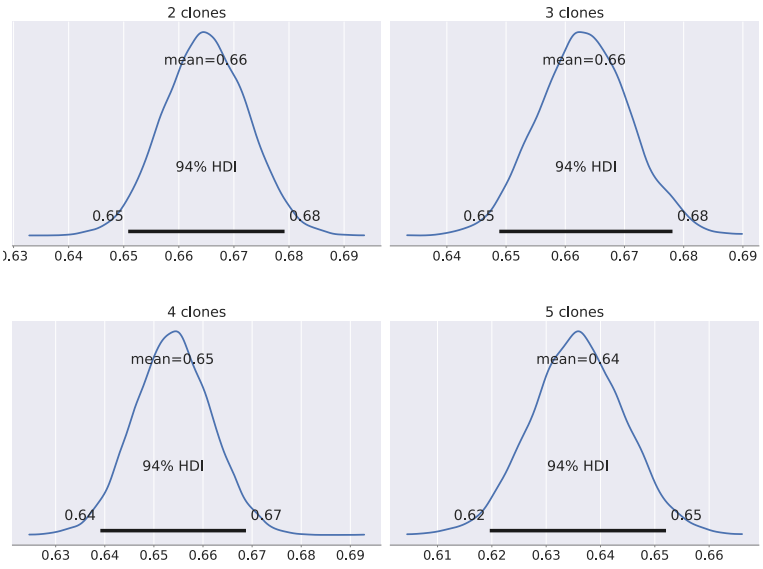


(a) Distributions of 94% HDI widths across ten replicates for the base and extended models at four tumour fractions and three read depths. Superimposed strip plots assist in delineating model types.

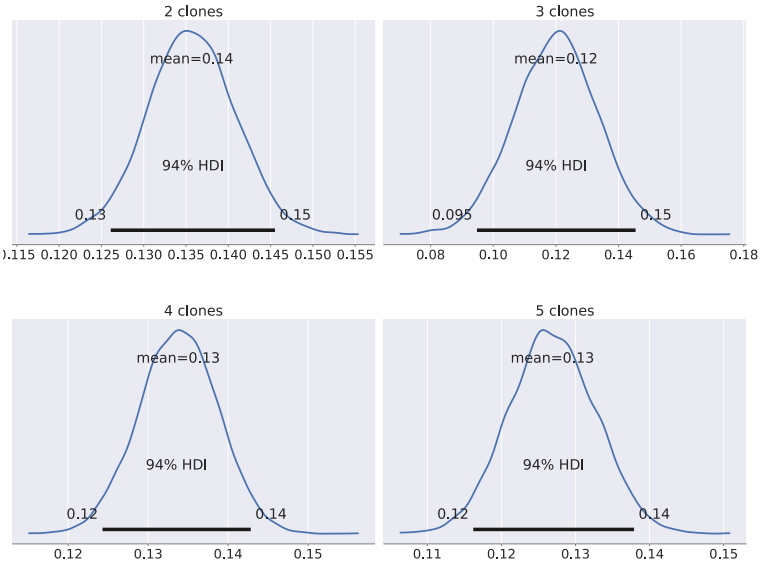


(b) Bar plots depicting the Bayesian coverage across ten replicates at four tumour fractions and three read depths.

Figure A.10: Posterior distribution statistics for semi-realistic experiments on read depth. **(a)** HDI widths **(b)** Bayesian coverage



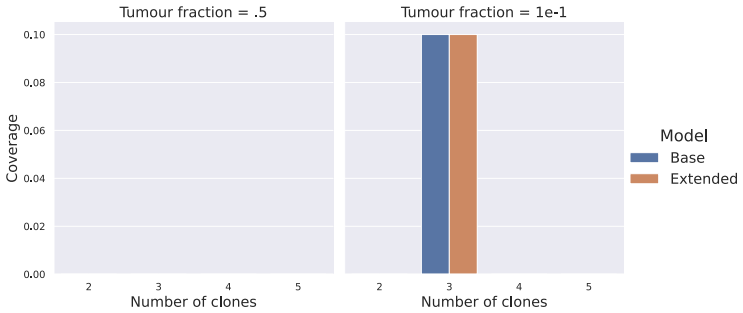
(a) Tumour fraction = .5.



(b) Tumour fraction = 1e-1.

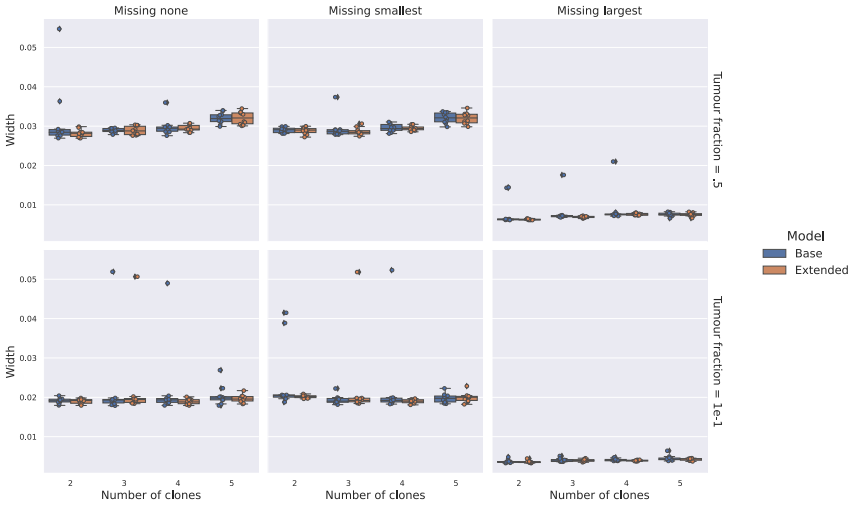
Figure A.11: Semi-realistic experiments posterior plots from both models for two to five clones. Dark black lines indicate the HDI.

(a) Distribution of 94% HDI widths across ten replicates for the base and extended models for two to six clones at two tumour fractions. Superimposed strip plots assist in delineating model types.

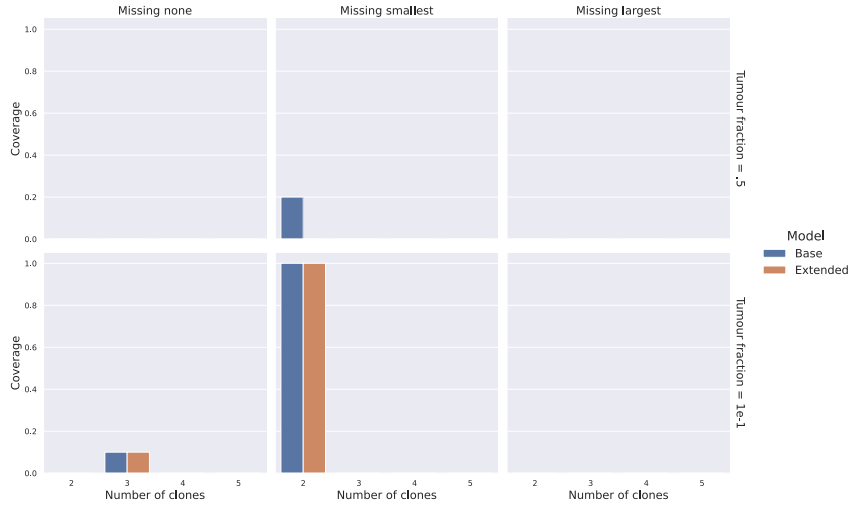


(b) Bar plot depicting the Bayesian coverage across ten replicates for the base and extended models for two to six clones at two tumour fractions.

Figure A.12: Posterior distribution statistics for semi-realistic experiments on number of clones. **(a)** HDI widths **(b)** Bayesian coverage



(a) Distributions of 94% HDI widths across ten replicates for two to five clones (includes the missing clone) at tumour fractions .5 and 1e-1 for semi-realistic experiments. A strip plot is superimposed for easily identifying model types.



(b) Bar plot depicting the Bayesian coverage across ten replicates for two to five clones (includes the missing clone) at two tumour fractions for semi-realistic experiments.

Figure A.13: Posterior distribution statistics for semi-realistic experiments on the effects of a missing clone. **(a)** HDI widths **(b)** Bayesian coverage