Classifying Copy Number Variations Using State Space Modeling of Targeted Sequencing Data: A Case Study in Thalassemia

Austin Talbot TALBOTA@PILLARBIOSCI.COM
Pillar Biosciences
Natick, MA, USA

Alex Kotlar AKOTLAR@EMORY.EDU
Department of Human Genetics
Emory University
Atlanta, GA, USA

Lavanya Rishishwar RISHISHWARL@PILLARBIOSCI.COM
Pillar Biosciences
Natick, MA, USA

Yue Ke KEY@PILLARBIOSCI.COM
Pillar Biosciences
Natick, MA, USA

Abstract

Thalassemia, a blood disorder and one of the most prevalent hereditary genetic disorders worldwide, is often caused by copy number variations (CNVs) in the hemoglobin genes. This disorder has incredible diversity, with a large number of distinct profiles corresponding to alterations of different regions in the genes. Correctly classifying an individual's profile is critical as it impacts treatment, prognosis, and genetic counseling. However, genetic classification is challenging due to the large number of profiles worldwide, and often requires a series of tests. Targeted next generation sequencing (NGS), which characterizes segments of an individual's genome, has the potential to dramatically reduce the cost of testing and increase accuracy. In this work, we introduce a probabilistic state space model for profiling thalassemia from targeted NGS data, which naturally characterizes the spatial ordering of the genes along the chromosome. We then use decision theory to choose the best profile among the different options. Due to our use of Bayesian methodology, we are also able to detect low-quality samples to be excluded from consideration, an important component of clinical screening. We evaluate our model on a dataset of 57 individuals, including both controls and cases with a variety of thalassemia profiles. Our model has a sensitivity of 0.99 and specificity of 0.93 for thalassemia detection, and accuracy of 91.5% for characterizing subtypes. Furthermore, the specificity and accuracy rise to 0.96 and 93.9% when low-quality samples are excluded using our automated quality control method. This approach outperforms alternative methods, particularly in specificity, and is broadly applicable to other disorders.

1. Introduction

Thalassemia is a genetic blood disorder that disrupts the production of hemoglobin, resulting in chronic anemia and other health complications Weatherall and Clegg (2001). It is one of the most prevalent genetic disorders worldwide and poses a significant public health challenge, particularly in regions such as the Mediterranean, Southeast Asia, and sub-Saharan Africa Weatherall and Clegg (2008); Modell and Darlison (2008). Thalassemia is caused by copy number variations (CNVs) or single nucleotide variants (SNVs) in the *HBA* (alpha-globin) and *HBB* (beta-globin) genes. In CNV-caused thalassemia, while individuals normally possess four copies of each gene, affected individuals have alterations in portions of one or more copies. The severity of anemia depends on the number of missing copies, ranging from mild symptoms to life-threatening conditions Galanello and Cao (2011); Cao and Galanello (2010); Rund and Rachmilewitz (2005). The portions of the genes that are altered can differ dramatically between individuals due to the diverse geographic origins of the disease. Accurate characterization of these "profiles" (specific patterns of alterations in the gene) is important for treatment and genetic counseling.

Traditional diagnostic methods rely on a sequence of hematologic assessments combined with genetic testing. Preliminary screening involves a complete blood count, which often reveals anemia with low mean corpuscular volume and an elevated red blood count. A peripheral blood smear may further support the diagnosis by showing microcytosis, hypochromia, target cells, and basophilic stippling. Hemoglobin electrophoresis or high-performance liquid chromatography (HPLC) is then used to quantify hemoglobin fractions. In β-thalassemia, elevated hemoglobin A2 (HbA2) and sometimes hemoglobin F (HbF) are observed, whereas α-thalassemia often results in normal HbA2 and HbF levels, making it more difficult to detect without molecular testing Ghosh et al. (2014); Adekile et al. (2015). Iron studies are typically conducted to rule out iron deficiency anemia (Galanello and Cao, 2011).

Next-generation sequencing (NGS) has transformed diagnosis of genetic disorders, allowing a small number of tests to completely characterize the genetic profile of the individual. In NGS, an individual's DNA is fragmented, amplified, and sequenced in parallel, allowing for the reconstruction of genomic sequences. While whole-genome sequencing (WGS) reconstructs the entire genome for a more complete characterization of disease status, targeted sequencing—where only specific regions of interest are amplified—provides a cost-effective alternative for disease diagnosis Rehm (2013); Ku and Roukos (2013). By designing amplicons that capture clinically relevant regions, targeted sequencing minimizes unnecessary data collection and optimizes throughput in clinical settings Mamanova et al. (2010); Rabbani et al. (2012).

Unfortunately, detecting CNVs, such as thalassemia, from NGS data remains a significant challenge Alkan et al. (2011). Unlike single-nucleotide variants (SNVs), which can be identified by sequence mismatches, CNV detection relies on quantifying read-depth differences relative to a reference sample (clinical normal). This process is highly susceptible to variability introduced by sequencing biases, amplification artifacts, and technical noise Teo et al. (2012). These difficulties are amplified in thalassemia profiling, which requires classification of the correct genetic profile of the thalassemia mutation. Existing CNV detection methods often require either (i) a large number clinical normal samples Krumm et al. (2012); Backenroth et al. (2014), which is impractical in clinical settings, or (ii) CNVs

spanning large genomic regions Talevich et al. (2016); D'Aurizio et al. (2016), which is not well suited for thalassemia where deletions can be small and localized. Moreover, current publicly-available approaches focus solely on identifying the existence of CNVs, the subsequent classification task is not included.

In this paper, we introduce a novel probabilistic approach for jointly estimating CNVs and classification of mutation types. While this is applicable to profiling a wide class of genetic disorders using targeted sequencing, we demonstrate its performance on thalassemia specifically. Our method is based on state-space modeling Chopin et al. (2020), a framework well-suited for sequential data, as amplicons are spatially ordered along the genome. This allows for more accurate copy number estimation by leveraging correlations between adjacent amplicons, while allowing for sudden transitions corresponding to deletions or amplifications. Using these estimates, we then apply Bayesian decision theory to assign each sample to a predefined thalassemia profile, accounting for both genetic variability and prior knowledge. Finally, we are able to evaluate sample quality via the Bayesian evidence of the state space model, quantifying the relative confidence in the model fit of the data. Our framework offers several advantages, namely (i) an estimation of copy number at each targeted region of the gene, refining CNV calls beyond simple presence/absence detection, (ii) a probabilistic classification of thalassemia profile, incorporating genetic inheritance patterns, and (iii) automatic sample quality evaluation using Bayesian evidence, allowing clinicians to assess confidence levels in each prediction, enhancing clinical interpretability. To properly implement these methods, we require altering standard inference of the state space model using sequential Monte Carlo in order to handle the abrupt changes in copy number. We evaluate the performance of our model on a mixture of synthetic and real data, and show that our method has high sensitivity and specificity, especially when (iii) is used to reject low-quality samples. A Python package implementing the models and code to reproduce synthetic results is available at https://github.com/Pillar-Biosciences-Inc/StateCNV.

Generalizable Insights about Machine Learning in the Context of Healthcare

- Bayesian evidence significantly enhances diagnostic accuracy by identifying low-quality samples that require manual review or rerunning.
- Auxiliary particle filtering improves performance in state space models, particularly in scenarios where abrupt changes are expected.

2. Related Work

There are three areas of related work; methods developed to detect thalassemia using NGS, methods for CNV quantification, and work with state space modeling of genomic data.

Traditional thalassemia profiling requires multiple assays: Sanger sequencing for SNVs and small indels, and GAP-PCR, MLPA Colosimo et al. (2011), or chromosomal microarrays for structural variants (SVs). While these methods have high sensitivity/specificity up to 0.99 depending on population ARUP Consult Editorial Board (2025), cost makes these methods inaccessible in many heavily-affected regions. NGS offers a comprehensive alternative, but early amplicon-based approaches targeted only specific SVs, requiring pooled PCR reactions. A more costly hybrid-capture NGS assay covering 275 kb of hemoglobin-related

regions was introduced by Shang et al. (2017). More recently, Cao et al. (2022) developed NGS4THAL, a bioinformatics tool for thalassemia variant detection, but it depends on whole-genome or hybrid-capture sequencing for continuous coverage. Despite these advances, clinical integration of NGS remains limited due to complex library preparation and the computational demands of SNV and SV detection.

The literature on more generic CNV estimation is quite extensive Gordeeva et al. (2021). Many of these methods, such as CNVkit Talevich et al. (2016) and Excavator Magi et al. (2013), are optimized for detecting large-scale copy number variations (CNVs) and typically require large sample cohorts for normalization. CNVkit Talevich et al. (2016) is designed for targeted and exome sequencing but relies on pooled normal samples or panel-based reference creation, making it less effective in small-cohort or single-sample settings. Excavator, developed for whole-exome and whole-genome sequencing, employs a hidden Markov model (HMM) to detect CNVs but assumes broad genomic alterations rather than small, localized deletions, such as those characteristic of thalassemia. Other tools, including GATK-gCNV Babadi et al. (2023) and CoNIFER Krumm et al. (2012), similarly depend on large reference panels to reduce noise and improve sensitivity, making them impractical for clinical applications where individual patient testing is required. Moreover, while these methods estimate CNVs, none provide a subsequent classification step that assigns the detected copy number changes to predefined disease-specific profiles. A more in-depth description of why many publicly-available CNV callers are unsuitable is provided in the appendix.

The use of state space models applied to genomics is quite extensive. Early work with hidden Markov models (HMMs) demonstrated their utility for large CNV detection (Scharpf et al., 2008; Xu et al., 2011). More general state-space approaches for genomic segmentation have been proposed Zhu et al. (2012). These models have substantial advantages over traditional CNV segmentation methods due to their explicit modeling of uncertainty and flexibility to adapt to customizable observational models. These models often depend on Markov chain Monte Carlo methods (specifically particle filtering) (Chopin et al., 2020). Among this class of algorithms, auxiliary particle filtering Douc et al. (2011) is the most pertinent to this work, as we use this technique to address difficulties associated with the abrupt changes occuring in the series.

3. Methods

3.1. Notation and Data

The data come from targeted amplicon sequencing. In this type of NGS, small sections of the DNA, known as amplicons, are repeatedly copied (amplified) using a process called PCR. After amplification, a subset of these copies are read using a sequencing machine and the number of times a particular amplicon is read is the read count. More copies of a gene in the original starting material will result in higher read counts, with noise due to both amplification and sequencing.

We analyze read counts from J targeted amplicons, of which $K = \{i_1, \ldots, i_K\}$ are control amplicons coming from regions known to be CNV neutral. The counts are obtained from the case sample using the Pivat® pipeline, denoted as $\{s_j\}_{j=1:J}$, and a corresponding normal sample, $\{n_j\}_{j=1:J}$. To assess copy number variations (CNVs), we compute the log copy ratio for each amplicon $\tilde{y}_j = \log(s_j + 10) - \log(n_j + 10)$ (the offset of 10 is used to

handle the case of homozygous deletions). Functionally, the purpose of the normal sample is to account for amplicon-specific effects on the read counts. Often, there are systematic biases in read counts between the sample and normal stemming from different sequencing depths. We use the control amplicons to account for this by subtracting the mean of the log ratios in the control amplicons, as

$$y_j = \tilde{y}_j - Mean(\tilde{y}_K). \tag{1}$$

After this transformation, $y_j = 0$ indicates a copy-neutral amplicon (identical to the control amplicons), while negative values indicate deletions and positive values indicate amplifications. We refer to these covariates as the log copy number ratio (lCNR), as these quantities ideally are the log of the ratio of copy numbers between the sample and the control. A visualization of this procedure is given in Figure 1.

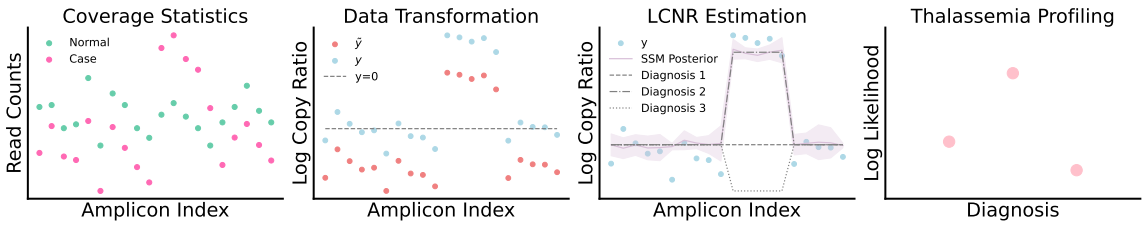

Figure 1: A visualization of our analysis procedure. The read counts are obtained from a sample and associated control. We then compute the log ratio of these counts, adjusting for relative abundance. We then estimate the ratio at each amplicon. Finally, we compare the estimates to predefined profiles (dotted lines) to classify subtype.

Thalassemia profiling relies on clinically predefined "profiles", each representing expected log ratios in regions in the genome (and by extension the amplicons). We denote the expected log ratio for the i-th profile as $d_i = \{d_{ij}\}_{j=1:J}$. These profiles vary based on population-specific mutations (i.e. Southeast Asian or Turkish deletions). These profiles may correspond to heterozygous or homozygous deletions and amplifications. For a given sample we must estimate the number of copies at each amplicon, and subsequently identify the most likely thalassemia profile for both the alpha and beta genes. Example profiles are shown in the top right of Figure 2.

3.2. A State Space Model of the Log Copy Ratios

The lCNRs exhibit strong spatial correlation due to the physical ordering of the amplicons along the chromosome. A natural model for this dependency is through a state-space model Chopin et al. (2020). According to these models, the observed lCNRs are noisy measurements of the true lCNR. Correlations in these unobserved values are induced via a Markov chain of the true lCNRs. Formally, denoting the unobserved true lCNR as x_j, we model this chain as

$$p(x_{j+1} \mid x_j) = (1-p)\mathcal{N}(x_j, \epsilon) + p\mathcal{N}(x_j, \sigma^2). \tag{2}$$

Here $p \approx 0.01$ can be viewed as the small probability of a CNV event (jump in copy number), while $\sigma^2 \gg \epsilon$ ensures that most transitions are small but CNVs lead to significant jumps. Although a piecewise-constant model (where $\epsilon = 0$) might be more biologically justifiable, the backwards algorithm described later requires a non-atomic distribution. Instead, we allow for small fluctuations in copy number estimates while preserving the ability to detect abrupt shifts. A visualization of random trajectories sampled from this process is given in the top left of Figure 2.

We model the noise in the observed lCNRs using a Laplace likelihood

$$p(y_j \mid x_j) = \frac{1}{2b} \exp\Big(-\frac{|y_j - x_j|}{b}\Big). \tag{3}$$

The Laplace distribution is heavier-tailed than the more common Gaussian distribution, preventing outliers due to sequencing artifacts and other sources of noise from overly influencing inference of the true lCNRs. Finally, we impose a weak prior on the first $lCNR$ of $p(x_0) = \mathcal{N}(0, \tau)$, corresponding to a weak assumption that it is copy-neutral. The complete state-space model is

$$\begin{aligned} p(x_0) &= \mathcal{N}(0, \tau), \\ p(x_{j+1} \mid x_j) &= (1-p)\mathcal{N}(x_j, \epsilon) + p\mathcal{N}(x_j, \sigma^2), \\ p(y_j \mid x_j) &= \text{Laplace}(x_j, b). \end{aligned} \tag{4}$$

The choices for the parameters p, σ^2, ϵ and b are chosen a priori. The choice of ϵ is not important, as it is set to be very small to represent the fact that the CNR is flat within regions. The parameter σ^2 is not sensitive, provided it is large enough to allow for large jumps. The choice of p and b are more important and based on previous runs. Good heuristics for their choice are that p can be set to the expected number of CNVs in a positive sample divided by the number of amplicons, and b initialized either from the spread of observed ratios or via likelihood maximization on normal samples. We provide an analysis of parameter choice on synthetic data in the appendix.

3.3. Inference Via Sequential Monte Carlo

The state-space model in (4) does not provide a simple closed-form expression for the posterior distribution of $p(x \mid y)$, the lCNR given the observed counts. A common solution in Bayesian inference is to use Monte Carlo methods Robert and Casella (2004), which generate samples from the posterior distribution to approximate key quantities such as means, medians, and quantiles. When direct sampling is intractable, a Markov chain can be constructed that converges to the posterior distribution. However, standard Markov Chain Monte Carlo (MCMC) methods Betancourt (2017) perform poorly for state-space models due to high correlations in the latent variables, making inference unreliable for all but the shortest sequences Chopin et al. (2020). Sequential Monte Carlo, or particle filtering, overcomes these limitations by using the Markov property of state-space models,where x_{j+1} depends only on x_j, to improve inference quality significantly. There are three components we discuss: the forward filtering algorithm, the backward sampling algorithm, and auxiliary filtering.

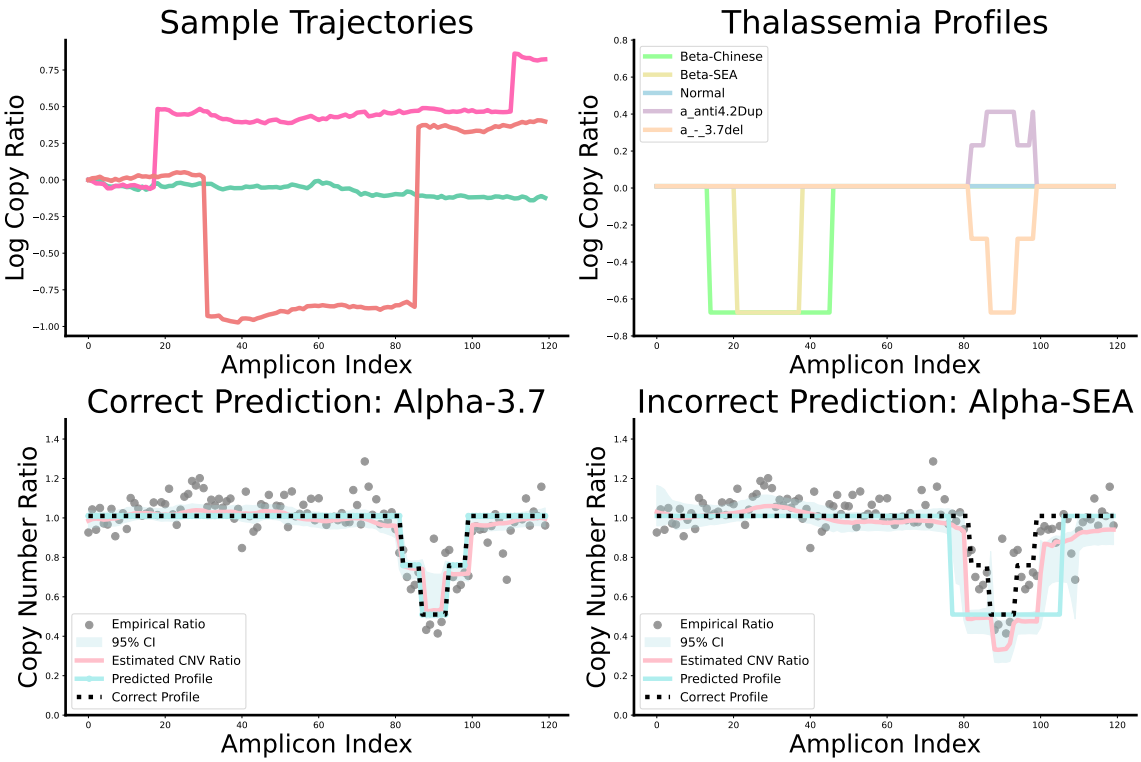

Figure 2: The top left shows sample trajectories drawn from 2 with $\sigma^2 = 1$, $p = 0.02$, and $\epsilon = .001$. The top right shows the profiles of 5 different thalassemia profiles, including two beta deletions, an alpha deletion, and an alpha duplication. The bottom left shows a sample predicted correctly by our model, including the posterior mean, credible interval, and diagnosis. The bottom right is identical, except the model predicted south-east Asian deletion (SEA) rather than the correct profile alpha 3.7 deletion.

3.3.1. Forward Filtering Algorithm

The forward-filtering step approximates $p(x_j \mid y_{0:j})$ using a custom form of importance sampling Goodfellow et al. (2016). The algorithm is

1. Initialization: Generate N particles from the prior distribution: $X_0^N \sim p(x_0)$.
2. Weighting: Assign importance weights based on how well the data align with each particle $w_0^N = p(y_0|X_0^N)$
3. Auto-normalize the weights: Make the average value of the weights sum to 1 as $W_0^N = w_0^N / \sum_{i=1}^N w_0^i$

4. Resampling: Select particles according to their weights, favoring those with higher likelihoods
$$A_0^{1:N} = \text{resample}(W_0^{1:N}) \tag{5}$$

5. Propogation: Generate new samples using the transition kernel:
$$X_1^{1:N} \sim p(x_1|X_0^{A_0^N}) \tag{6}$$

6. Iteration: Repeat this process for all time steps J, producing a sequence of weighted samples (X_j^N, W_j^N) that approximate each posterior distribution.

Like importance sampling, any quantity can be computed as a weighted sum of the function applied to the individual samples. However, unlike standard importance sampling it incorporates a resampling step. Traditional importance sampling estimates quickly become effectively based on a single "best" sample. While resampling increases the variance of the estimate in the short term, by "pruning" the bad samples it allows for multiple good samples to be propagated to the next element in the sequence.

3.3.2. Backward Smoothing for Posterior Estimation

The forward algorithm allows us to characterize $p(x_j \mid y_{0:j})$, which means the estimates only use information in the spatially preceding amplicons. Ideally, we would like the estimates to use the information of all amplicons in our estimate of the lCNR, particularly for the initial amplicons.This distribution, $p(x_j \mid y_{0:J})$, is significantly more challenging and requires an additional step called backward smoothing to approximate. Intuitively, backwards sampling starts at the end of the sequence and tracks where the origins of the final samples recursively. The algorithm for a single trajectory is

1. Initialize at the final time step J. The posterior $p(x_J|y_{0:J})$ is approximated by the forward filter weights W_J^N.

2. Sample the final point as $B_J \sim \text{Mult}(W_J^{1:N})$.

3. Compute the backward weights $\hat{w}_{J-1}^n = W_{J-1}^n p_J(X_J^{B_J}|X_{J-1}^n)$

4. Auto-normalize the weights $\hat{W}_{J-1}^n = \hat{w}_{J-1}^n / \sum_{m=1}^N \hat{w}_{J-1}^m$ for all n

5. Sample $B_{J-1} \sim \text{Mult}(\hat{W}_{J-1}^n)$

6. Iterate until $T = 0$.

From this we obtain a single simulated trajectory $(X_0^{B_0}, \ldots, X_J^{B_J})$. To approximate the posterior we generate M trajectories, giving a total computational complexity of $\mathcal{O}(JNM)$. We use the common choice $M = N$, and once these N trajectories are obtained we can obtain an estimate of the posterior mean as
$$lCNR_j = \frac{1}{M} \sum_{m=1}^{M} X_j^{B_j^m}. \tag{7}$$

3.3.3. Guided Particle Filtering

The propagation step in the forward filter using the transition kernel is the most straightforward approach (known as the bootstrap filter). However, it performs poorly with this particular transition kernel. Given that we want to the parameter p to be small (CNVs are infrequent), most of the samples for x_{j+1} will be close to the current value x_j. When a transition occurs at the edge of the CNV, most of these samples will be far away from the new value, and as a result inference will effectively be based on a small number of samples.

To address this issue, we use guided particle filtering (also known as auxiliary particle filtering, APF). This approach introduces a proposal distribution that selectively generates particles in regions more likely to match the observations. Instead of directly sampling from the transition model, we introduce an alternative proposal density, denoted as $q(x_j \mid x_{j-1})$, with more desirable properities. As long as this altered proposal distribution is "inverted" by changing how the weights are computed, inference remains unaltered.

In this work, we chose a transition density that generates the majority of the particles near the previous lCNR estimate x_{j-1}. However, we also generate a small proportion of samples near the current empirical lCNR y_j. This means that when an abrupt change does occur, there will be some samples near the new value. The guide we used that incorporates these assumptions is a Gaussian mixture model

$$q(x_j \mid x_{j-1}, y_j) = 0.8\mathcal{N}(x_{j-1}, \sigma^2) + 0.2\mathcal{N}(y_j, \sigma^2). \tag{8}$$

The weights were chosen heuristically such that most samples were likely to be at the current value but sufficient samples would properly explore regions associated with jumps.

3.4. Thalassemia Profiling via Bayesian Decision Theory

So far, our focus has been on the first task of estimating the lCNR at each amplicon. However, our ultimate goal is to assign both an alpha thalassemia and beta thalassemia profile to the sample. To accomplish this, we develop a probabilistic method for determining the profile, denoted D, by viewing the predefined profiles as the expected value of the lCNR, that is $d_{ij} = E[x_j \mid D = i]$. We assume that the observational likelihood is Gaussian, that is $p(x_j \mid D = i) = \mathcal{N}(d_{ij}, \sigma^2)$, where σ^2 is a tuning parameter controlling how informative a single amplicon is. Large values of σ^2 correspond to low confidence in any particular amplicon for determining the profile, small values instead view each amplicon as highly informative. In this portion of the classifier, we will ignore the spatial dependence of the amplicons. Given that σ^2 is a tuning parameter controlling the information provided by a single amplicon, full modeling of the spatial correlations is redundant.

Under these assumptions, if the true copy numbers were known, the optimal profile assignment could be determined using Bayesian decision theory as

$$\hat{D} = \arg\max_i \prod_{j=1}^{J} p(x_j \mid D = i) p(D = i). \tag{9}$$

However, we do not observe x_j directly; we only obtain posterior samples given the empirical log ratios. However, we can integrate out the true copy ratios as

$$p(D = i \mid y_{1:J}) = \int p(D = i \mid x_{1:J}) p(x_{1:J} \mid y_{1:J}) dx_{1:J}. \tag{10}$$

This distribution can be approximated using the samples provided by the backward smoothing step. The final profile assignment is then simply selecting the most probable profile. A visualization of the output of our aggregate caller is shown in the bottom row of Figure 2.

3.5. Sample Quality Evaluation via Model Evidence

Finally, a key concern in many clinical applications is determining whether a sample meets the quality standards necessary for making a reliable diagnosis. It is important to note that this is distinct from having an "uncertain" high-variance posterior. Intuitively, an ambiguous opinion after observing high-quality data is very different from ambiguity stemming from low quality data. Any probabilistic model, such as ours defined in (4), provides a natural way to assess the quality of the data as measured by how "realistic" the observed data are according to the generative model via the normalization constant of the model. This constant, also known as the Bayesian evidence, can be used for this purpose in any Bayesian model. In spite of its generality, it is rarely used as reliably estimating the Bayesian evidence is difficult. Fortunately, the particle filter offers a straightforward estimator: the mean of the unnormalized importance weights, given by $\ell_j = \frac{1}{N}\sum_{i=1}^{N} w_i$.

The total normalization constant for a sample is then computed as the product of these individual normalization constants $L = \prod_{j=0}^{J} \ell_j$. While L provides a measure of model fit, its absolute value is arbitrary and depends on the units of the parameters. Additionally, simple parameter choices, such as b or σ^2, can significantly influence these values. However, L is useful for comparing samples relative to other samples. Low values compared to other samples indicate that a sample is less compatiable with the model, while high values indicate a better model fit. This provides a natural quality control mechanism—samples with low likelihood scores should be flagged and returned to the clinician for manual review.

As previously noted, high-quality samples containing homozygous deletions can violate model assumptions (namely homozygous deletions leading to 0 counts), leading to artificially low values of L. To mitigate this issue, we use a more robust metric: the median of the individual likelihood values

$$QC_i = \text{Median}(\ell_{j\,j=0:J}). \tag{11}$$

Based on the properties of the median, this will ignore the subset of values violating model assumptions Huber and Ronchetti (2009), allowing us to evaluate sample quality based on a typical amplicon. We set a threshold for rejecting samples: those with a QC_i score more than three standard deviations below the mean are considered low quality and excluded from further analysis.

4. Cohort

4.1. Data Processing

We first performed read-to-genome alignment to the GRCh37/hg19 genome using BWA-MEM (Li, 2013). The initial alignment is improved by performing local realignment using Smith-Waterman (Smith et al., 1981) and a proprietary algorithm. To maximize the base accuracy and minimize the sequencing noise, paired-end reads are assembled into consensus reads, weighted with the base quality scores from both mates. The assembled reads correspond to the gene-specific positions on the genome. A set of filters are applied to remove

non-uniquely mapped reads (e.g., pseudogenes) and reads that do not match the amplicon positions (e.g., primer-dimers or non-specific amplifications). For CNV calculations, per-amplicon read depth is calculated for each sample. Counts of overlapping amplicons are stored as independent entries.

4.2. Thalassemia Samples

We obtained 44 positive and 13 negative clinical samples as de-identified remnants from a laboratory that tested the samples for clinical use and would have otherwise discarded them. FDA guidance states that such samples may be used for research purposes without obtaining patient consent. These samples were sequenced using a panel of 131 amplicons that covers alpha, beta, delta, gamma and epsilon Thalassemia regions, as well as pseudogenes and control regions. Sequencing was performed on an Illumina MiSeq^{TM} platform to an average depth of approximately 2800 paired-end reads. The set of true alpha thalassemia diagnoses were obtained from an orthogonal, clinically validated test that uses MLPA with reflex to GAP-PCR for alpha-thalassemia deletion and duplication detection. Of the positive samples, 31 had a heterozygous alpha-3.7 deletion, 5 had a homozygous alpha-3.7 deletion, 2 had alpha-4.2 deletions, 3 had south-east Asian deletions, 1 sample with an alpha-4.2 duplication, 1 sample with a large HBA duplication, and 1 sample with an alpha-3.7 deletion/4.2 duplication.

4.3. BRCA Samples

To further investigate our quality control metrics, we obtained a second dataset sequenced using a panel designed to detect mutations in the BRCA1 and BRCA2 genes. This panel contains 283 amplicons targeting various regions within BRCA1 and BRCA2 genes, as well as 8 control amplicons from other chromosomes. We obtained four mutation-positive DNA cell line samples from the Coriell Institute for Medical Research with known mutations: NA18949 with a BRCA1 exon 14 and 15 deletion, NA14626 with a BRCA1 exon 12 duplication, NA0330 with a whole-gene duplication of BRCA2, and NA02718 with a whole gene deletion of BRCA2. We also obtained four negative cell line samples NA12878, NA19240, NA24385, and NA24143. Additionally, Horizon Discovery's Mimix^{TM} Quantitative Multiplex, fcDNA (Moderate) Reference Standard was used to evalue performance on Formalin-compromised DNA (FFPE). Libraries were prepared and sequenced on Illumina's MiSeq^{TM}. All samples were collected under informed consent.

5. Results

5.1. Evaluating the Auxiliary Filter in Synthetic Data

In this section, we focus on evaluating the impact of our guided filter as compared to the standard bootstrap filter. Traditionally, the effective sample size (ESS), which estimates the number of "independent" samples generated by the MCMC algorithm, is commonly used to evaluate performance. However, we introduce additional metrics, namely the root mean squared error (RMSE) of the CNV estimates and the estimated Bayesian evidence. These provide a more comprehensive comparison between the two inference algorithms, as they measure the quantities produced by our model.

To generate synthetic data, we simulated a panel with 160 amplicons containing two CNVs: one homozygous duplication and one heterozygous deletion, along with 40 control amplicons. We then fit two models: one using the standard bootstrap filter and the other employing the guided particle filter. Both models used 500 particles with identical state-space model parameters and the metrics were evaluated on each. Each experiment was repeated 10 times to compute confidence intervals for all metrics.

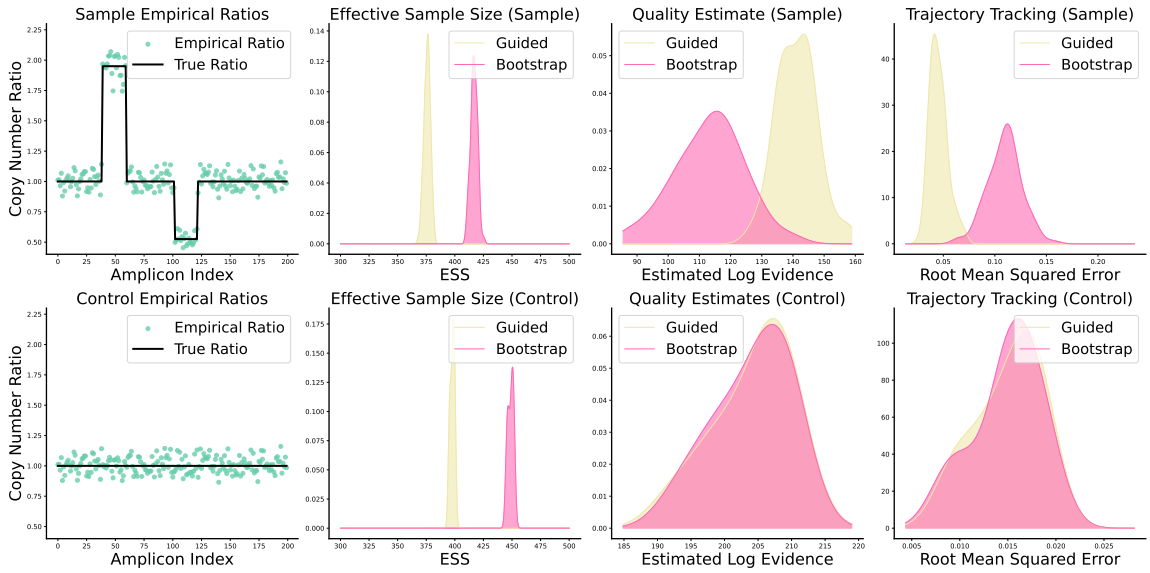

Figure 3: The top left shows the empirical CNR from a single synthetic sample with two CNVs, along with the true CNR. The plot in the middle left shows the distribution effective sample size of guided filter in yellow and bootstrap filter in red. Middle right shows the distribution of the log evidence of the sample. Far right shows the RMSE of the posterior median predicting the CNR. The bottom row repeats this process but with a synthetic sample lacking CNVs.

The results for a representative trial are visualized in Figure 3. As expected, the guided filter produces a lower effective sample size of 375 ± 5 compared to 417 ± 5 for the bootstrap filter. Despite this, it substantially improves model fit, particularly in detecting transition points. This is reflected in the likelihood scores: 141.4 ± 12.6 for the guided filter versus 113.4 ± 21.8 for the bootstrap filter. The improved fit is further supported by a significantly lower RMSE (0.045 ± 0.017) for the guided filter compared to 0.109 ± 0.028 for the bootstrap filter.

To ensure that the guided filter's ability to track abrupt changes in copy number did not degrade performance on normal samples lacking CNVs, we also analyzed a synthetic "normal" (bottom row of Figure 3). The guided filter again showed lower sample size with 398 ± 3 as opposed to 449 ± 4 for the boostrap filter. However, the remaining metrics were largely unaltered; evidence estimates for the guided filter were (204.1 ± 10.8) as opposed to 204.1 ± 10.4 for the bootstrap, and the rMSE values were similarly close 0.0155 ± 0.006 versus

0.0149 ± 0.006. While we expect the bootstrap filter to yield marginally higher evidence due to its increased number of samples centered at the current value, the guided sampler does not introduce any statistically significant degradation in performance. Beyond the validation of our inference algorithm, these results illustrate that ESS is not always the most pertinent metric of performance in MCMC.

5.2. Evaluating the Thalassemia Profiler and Sample Quality Metric

There are two overall objectives in this section; first, we wish to evaluate the accuracy of our thalassemia profiling and second, we wish to evaluate the ability of Bayesian evidence to recognize substandard samples. To evaluate profiling accuracy, we use the thalassemia dataset from Section 4. To our knowledge, we have no competitor methods to evaluate the performance of our entire informatics pipeline. Instead, we compare the state space portion of the method against other standard smoothing methods. To evaluate the efficacy of Bayesian evidence for quality control, we use data from the BRCA panel. The process of fixing DNA with formalin when generating FFPE samples causes DNA degradation, leading to erratic empirical copy number ratios when an unmatched (non-FFPE) cell line normal is used as a reference for an FFPE sample. Thus, the FFPE samples in this dataset serve as known "poor-quality" samples, allowing us to evaluate the quality score's effectiveness in distinguishing between cell line and FFPE samples.

5.2.1. Evaluating the Profiler and Quality Control in Real Data

We first evaluate our ability to both detect thalassemia as well as classify thalassemia profiles. To measure this, we evaluated both the sensitivity and specificity of a positive test as opposed to negative, and the accuracy of the profile choice. We were able to increase sample size by separately considering diagnoses of the alpha and beta status, and by repeating the experiment with three different sets of normal samples for comparison. With this procedure, we obtained 324 measurements, of which 132 were positive and 192 were negative. We were unable to find any other publicly available profilers to provide a full bioinformatic comparison of our method. However, were able to compare our state space model to other methods for denoising the copy number estimates, providing a comparison of the first component of our profiler. We chose comparisons of no smoothing (empirical), kernel smoothing (Wasserman, 2006), smoothing splines (Hastie et al., 2009), and anisotropic diffusion (Black et al., 1998). We compare our method both when all samples are included (no QC) as well as the performance when subpar samples are removed (QC) according to the procedure described in Section 3.5. The results are shown in Table 1.

We found that even without quality control, our method had the highest accuracy of all considered methods with an accuracy of 91.5% as compared to the next highest (empirical with 91.2%). This further rises to 93.9% when substandard samples are excluded. Our method has lower specificity than smoothing splines 0.93 as opposed to 0.94, but this is rectified by quality control, which raises the specificity to 0.96, the best among all considered methods.

We also evaluate the ability of our model to estimate the CNR on an amplicon-by-amplicon basis, as specific amplicons or exons may be of interest in addition to predefined profiles. We can reuse the experimental setup from the previous section, except now we

Model	Accuracy	TP	FP	FN	TN	Sens.	Spec.	RMSE
Empirical	91.2%	131	20	1	172	0.99	0.90	0.156
Kernel Smoothing	90.9%	131	19	1	173	0.99	0.90	0.133
Smoothing Splines	90.4%	129	12	3	180	0.98	0.94	0.135
Anisotropic Diffusion	90.1%	131	19	1	173	0.99	0.90	0.131
Ours (No QC)	91.5%	131	14	1	181	0.99	0.93	0.131
Ours (QC)	**93.9%**	128	8	1	178	0.99	**0.96**	**0.111**

Table 1: Comparison of predictive performance of the state space model, both with quality control on the quality score (QC) and without such filtering (No QC), as compared to other smoothing methods

will compare the output of the state space model to evaluate how close the CNR is to the true ratio. We found that without QC our method matched the root mean squared error (rMSE) of the best smoothing method (anisotropic diffusion). However, once low-quality samples were excluded, the RMSE of our method was superior with a value of 0.111.

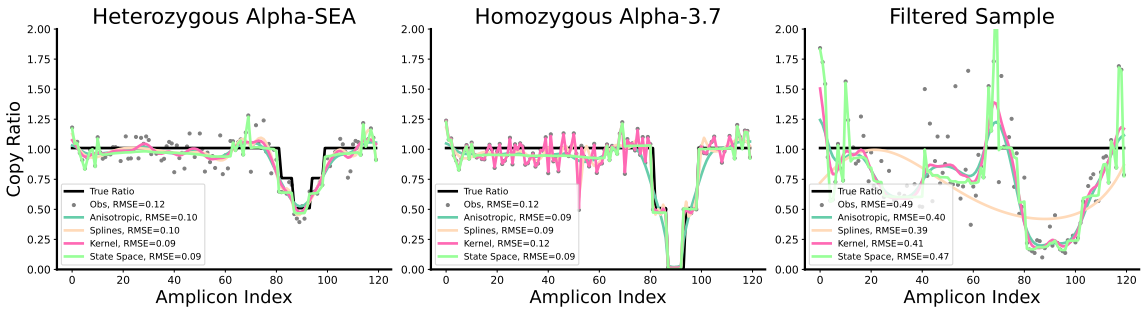

Figure 4: The results of the different smoothing methods on three different samples. The third plot shows a sample that was rejected by our quality control criterion.

5.2.2. Evaluating the Sample Quality Metric

Finally, we evaluate the ability of our model to detect poor-quality samples for quality control. While we already have evidence for its efficacy due to it improving specificity in Section 5.2.1, here we quantify its performance in differentiating low-quality samples from higher quality samples. We fit our state space model on each sample compared to three of the Coriell normals, evaluated the log evidence, and then calculated each sample's quality score. The results are shown in Figure 5. On the left, we show the ratios associated with one of the FFPE samples. We can see that the empirical copy ratios are far more erratic in the FFPE samples as compared to the clinical sample shown in the center. This is reflected in the quality scores obtained; the average score for the TV samples and clinical samples are

0.42 ± 0.08 and 0.29 ± 0.08 respectively while the scores for the FFPE samples are -2.66 ± 0.8. From this, we can see a clear separation between the a priori "low-quality" samples and the samples that match model assumptions. Based on the quality metric introduced in the methods, all samples from FFPE would have failed quality control and failed to produce a call. By contrast, only a single sample from the cell line positive would have been falsely rejected as being substandard.

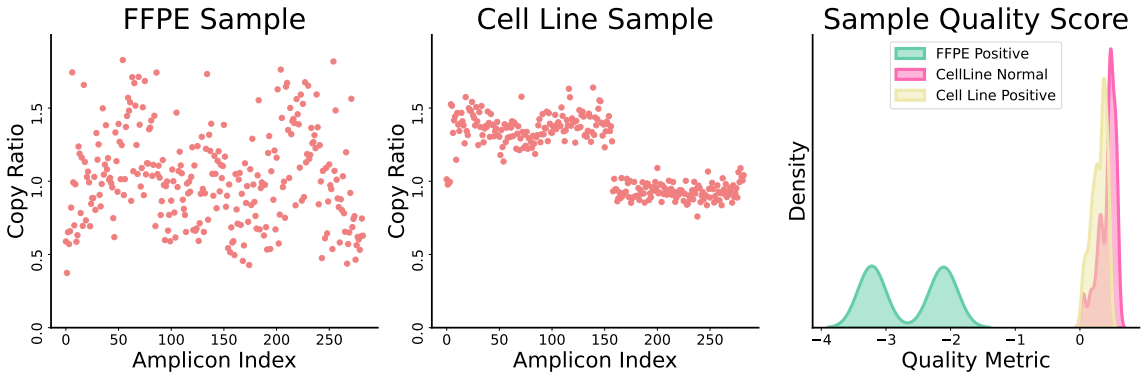

Figure 5: This figure visualizes the sample quality differences detected by our metric. On the left, we show the CNRs of an FFPE sample with a cell-line normal. These ratios fluctuate wildly between 0.5 and 1.5, with a standard deviation of 0.6. In comparison, the ratio of a cell line positives with the same normal shown in the middle plot has substantially reduced noise in the CNRs, with a standard deviation of 0.16. Finally, on the right we visualize the QC score standardized by batch to remove run-specific variability. We can see that the FFPE positive samples are all below -2.0, while both positive and normal cell-line samples were above 0.

6. Discussion

From a clinical perspective, this study makes two key contributions. First, to our knowledge, this is the first open-source thalassemia profiler using targeted sequencing. Furthermore, given the widespread prevalence of CNV disorders, this tool has significant clinical applications for other disorders beyond thalassemia, including the BRCA variations mentioned in 4. Second, our work demonstrates the utility of Bayesian evidence in identifying low-quality samples, enabling clinicians or technicians to flag them for manual review or rerunning.

From a machine learning perspective, our work highlights the importance of auxiliary particle filtering in detecting and characterizing change points using state-space models. In regions with dramatic shifts, the standard bootstrap filter tends to under-sample critical areas, leading to reduced predictive accuracy and an underestimate of the Bayesian evidence in the sample. Traditional metrics, such as the average effective sample size, fail to capture this issue and may misleadingly suggest the superiority of the bootstrap filter.

Limitations While our method performs well, there are several remaining limitations. First, our evaluation was limited both in sample size and variety in profiles. A larger cohort, including more obscure variants, would be highly beneficial in improving parameter choices and performance evaluation. Second, our method currently requires incorporation of an in-batch normal, which can increase costs of diagnosis. Future work will explore alternative strategies, such as direct modeling of amplicon efficiency, allowing for true normal-free calling, or alternatively incorporation of a "panel of normals" strategy designed to repeatedly reuse large numbers of normals in new runs.

References

Adekunle D Adekile, Asma F Azab, Sondus I Al-Sharida, Bahia A Al-Nafisi, Nagihan Akbulut, Rajaa A Marouf, and Nada Y Mustafa. Clinical and molecular characteristics of non-transfusion-dependent thalassemia in kuwait. *Hemoglobin*, 39(5):320–326, 2015.

Can Alkan, Bradley P Coe, and Evan E Eichler. Genome structural variation discovery and genotyping. *Nature reviews genetics*, 12(5):363–376, 2011.

ARUP Consult Editorial Board. Alpha thalassemia — test fact sheet. https://arupconsult.com/ati/alpha-thalassemia, 2025. Last Literature Review: January 2021; Last Update: May 2025.

Mehrtash Babadi, Jack M Fu, Samuel K Lee, Andrey N Smirnov, Laura D Gauthier, Mark Walker, David I Benjamin, Xuefang Zhao, Konrad J Karczewski, Isaac Wong, et al. Gatk-gcnv enables the discovery of rare copy number variants from exome sequencing data. *Nature genetics*, 55(9):1589–1597, 2023.

Daniel Backenroth, Jason Homsy, Laura R Murillo, Joe Glessner, Edwin Lin, Martina Brueckner, Richard Lifton, Elizabeth Goldmuntz, Wendy K Chung, and Yufeng Shen. Canoes: detecting rare copy number variants from whole exome sequencing data. *Nucleic acids research*, 42(12):e97–e97, 2014.

Yoshua Bengio, Ian Goodfellow, and Aaron Courville. *Deep learning*, volume 1. MIT press Cambridge, MA, USA, 2017.

Michael Betancourt. A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434*, 2017.

Michael J Black, Guillermo Sapiro, David H Marimont, and David Heeger. Robust anisotropic diffusion. *IEEE Transactions on image processing*, 7(3):421–432, 1998.

Antonio Cao and Renzo Galanello. Beta-thalassemia. *Genetics in medicine*, 12(2):61–76, 2010.

Yujie Cao, Shau-yin Ha, Chi-Chiu So, Ming-for Tony Tong, Clara Sze-man Tang, Huoru Zhang, Rui Liang, Jing Yang, Brian Hon-Yin Chung, Godfrey Chi-Fung Chan, et al. Ngs4thal, a one-stop molecular diagnosis and carrier screening tool for thalassemia and other hemoglobinopathies by next-generation sequencing. *The Journal of Molecular Diagnostics*, 24(10):1089–1099, 2022.

Nicolas Chopin, Omiros Papaspiliopoulos, et al. *An introduction to sequential Monte Carlo*. Springer, 2020.

Alessia Colosimo, Valentina Gatta, Valentina Guida, Eleonora Leodori, Enrica Foglietta, Silvana Rinaldi, Maria Pia Cappabianca, Antonio Amato, Liborio Stuppia, and Bruno Dallapiccola. Application of mlpa assay to characterize unsolved α-globin gene rearrangements. *Blood Cells, Molecules, and Diseases*, 46(2):139–144, 2011.

Romina D'Aurizio, Tommaso Pippucci, Lorenzo Tattini, Betti Giusti, Marco Pellegrini, and Alberto Magi. Enhanced copy number variants detection from whole-exome sequencing data using excavator2. *Nucleic acids research*, 44(20):e154–e154, 2016.

Randal Douc, Aurélien Garivier, Éric Moulines, and Jimmy Olsson. Sequential monte carlo smoothing for general state space hidden markov models. *The Annals of Applied Probability*, 21(6):2109–2145, 2011.

Peter I Frazier. A tutorial on bayesian optimization. *arXiv preprint arXiv:1807.02811*, 2018.

Renzo Galanello and Antonio Cao. Alpha-thalassemia. *Genetics in medicine*, 13(2):83–88, 2011.

Kanjaksha Ghosh, Roshan Colah, Mamta Manglani, Ved Prakash Choudhry, Ishwar Verma, Nishi Madan, Renu Saxena, Dipty Jain, Neelam Marwaha, Reena Das, et al. Guidelines for screening, diagnosis and management of hemoglobinopathies. *Indian journal of human genetics*, 20(2):101, 2014.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*. MIT press Cambridge, 2016.

Veronika Gordeeva, Elena Sharova, Konstantin Babalyan, Rinat Sultanov, Vadim M Govorun, and Georgij Arapidi. Benchmarking germline cnv calling tools from exome sequencing data. *Scientific reports*, 11(1):14416, 2021.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, 2nd edition, 2009. ISBN 978-0-387-84857-0.

Peter J Huber and Elvezio M Ronchetti. *Robust statistics*. John Wiley & Sons Hoboken, NJ, USA, 2009.

Jeffrey Humpherys and Tyler J Jarvis. *Foundations of Applied Mathematics Volume 2: Algorithms, Approximation, Optimization*. SIAM, 2020.

Niklas Krumm, Peter H Sudmant, Arthur Ko, Brian J O'Roak, Maika Malig, Bradley P Coe, Aaron R Quinlan, Deborah A Nickerson, Evan E Eichler, NHLBI Exome Sequencing Project, et al. Copy number variation detection and genotyping from exome sequence data. *Genome research*, 22(8):1525–1532, 2012.

Chee-Seng Ku and Dimitrios H Roukos. From next-generation sequencing to nanopore sequencing technology: paving the way to personalized genomic medicine. *Expert review of medical devices*, 10(1):1–6, 2013.

Heng Li. Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. *arXiv preprint arXiv:1303.3997*, 2013.

Alberto Magi, Lorenzo Tattini, Ingrid Cifola, Romina D'Aurizio, Matteo Benelli, Eleonora Mangano, Cristina Battaglia, Elena Bonora, Ants Kurg, Marco Seri, et al. Excavator: detecting copy number variants from whole-exome sequencing data. *Genome biology*, 14: 1–18, 2013.

Lira Mamanova, Alison J Coffey, Carol E Scott, Iwanka Kozarewa, Emily H Turner, Akash Kumar, Eleanor Howard, Jay Shendure, and Daniel J Turner. Target-enrichment strategies for next-generation sequencing. *Nature methods*, 7(2):111–118, 2010.

Bernadette Modell and Matthew Darlison. Global epidemiology of haemoglobin disorders and derived service indicators. *Bulletin of the World Health Organization*, 86(6):480–487, 2008.

Bahareh Rabbani, Nejat Mahdieh, Kazuyoshi Hosomichi, Hirofumi Nakaoka, and Ituro Inoue. Next-generation sequencing: impact of exome sequencing in characterizing mendelian disorders. *Journal of human genetics*, 57(10):621–632, 2012.

Heidi L Rehm. Disease-targeted sequencing: a cornerstone in the clinic. *Nature reviews genetics*, 14(4):295–300, 2013.

CP Robert and G Casella. *Monte Carlo statistical methods*. Springer Location New York, NY, 2004.

Deborah Rund and Eliezer Rachmilewitz. β-thalassemia. *New England Journal of Medicine*, 353(11):1135–1146, 2005.

Robert B Scharpf, Giovanni Parmigiani, Jonathan Pevsner, and Ingo Ruczinski. Hidden markov models for the assessment of chromosomal alterations using high-throughput snp arrays. *The annals of applied statistics*, 2(2):687, 2008.

Xuan Shang, Zhiyu Peng, Yuhua Ye, Xinhua Zhang, Yan Chen, Baosheng Zhu, Wangwei Cai, Shaoke Chen, Ren Cai, Xiaoling Guo, et al. Rapid targeted next-generation sequencing platform for molecular screening and clinical genotyping in subjects with hemoglobinopathies. *EBioMedicine*, 23:150–159, 2017.

Temple F Smith, Michael S Waterman, et al. Identification of common molecular subsequences. *Journal of molecular biology*, 147(1):195–197, 1981.

Eric Talevich, A Hunter Shain, Thomas Botton, and Boris C Bastian. Cnvkit: genome-wide copy number detection and visualization from targeted dna sequencing. *PLoS computational biology*, 12(4):e1004873, 2016.

Shu Mei Teo, Yudi Pawitan, Chee Seng Ku, Kee Seng Chia, and Agus Salim. Statistical challenges associated with detecting copy number variations with next-generation sequencing. *Bioinformatics*, 28(21):2711–2718, 2012.

Larry Wasserman. *All of nonparametric statistics*. Springer Science & Business Media, 2006.

David J Weatherall and John B Clegg. Inherited haemoglobin disorders: an increasing global health problem. *Bulletin of the World Health Organization*, 79(8):704–712, 2001.

David J Weatherall and John B Clegg. *The thalassaemia syndromes*. John Wiley & Sons, 2008.

Yaji Xu, Bo Peng, Yunxin Fu, and Christopher I Amos. Genome-wide algorithm for detecting cnv associations with diseases. *BMC bioinformatics*, 12:1–10, 2011.

Bin Zhu, Jeremy MG Taylor, and Peter X-K Song. Signal extraction and breakpoint identification for array cgh data using robust state space model. *arXiv preprint arXiv:1201.5169*, 2012.

Appendix

Alternative CNV Models

There is an extensive body of work on detecting copy number variants (CNVs) from next-generation sequencing (NGS) data. However, none of the methods we identified adequately address the clinical requirements of targeted sequencing panels. Publicly available CNV detection tools generally exhibit one of two disqualifying characteristics: (i) they require large cohorts of clinical samples for reliable inference, or (ii) they assume that CNVs span long genomic regions (typically tens of kilobases). Both assumptions conflict with the design and clinical use of targeted liquid biopsy panels, which typically operate on small sample batches and must detect short CNVs with high precision.

Tool	PMID Reference	Year	Algorithm	Spatial Context	Small Sample	CN Estimate	Gains and Losses
Canoes	24771342	2014	GC, HMM	N	N	Y	Y
CLAMMS	26382196	2015	GC, HMM	Y	Y	N	Y
cn.MOPS	22302147	2012	GC, Mixture	N	N	Y	Y
CNVkit	27100738	2016	GC, median	N	Y	Y	Y
CODEX	25618849	2015	Segmentation	N	Y	Y	Y
CoNIFER	22585873	2012	SVD	N	N	Y	Y
EXCAVATOR2	27507884	2016	GC, Ratio	N	Y	Y	Y
ExomeDepth	22942019	2012	HMM	N	Y	Y	Y
ExonDel	25322818	2014	GC, Ratio	Y	Y	Y	N
FishingCNV	23539306	2013	SVD, Ratio	N	N	Y	Y
HMZDelFinder	34046589	2017	Thresholding	N	Y	Y	N
XHMM	23040492	2012	SVD, HMM	N	N	Y	Y

Table 2: Comparison of CNV detection tools for targeted/exome sequencing

Methods designed for whole-exome or whole-genome sequencing, such as cn.MOPS, CNVkit, and EXCAVATOR2, are optimized for extremely low read depth, often with a maximum of 100 reads per locus. In contrast, amplicon-based targeted panels frequently achieve minimum depths on the order of hundreds to thousands of reads. Low-depth methods suffer from high per-locus variance in copy number estimates, which they mitigate by assuming that CNVs are sufficiently long (>20 kb) so that aggregating read depth over large regions reduces noise and improves accuracy. This assumption is incompatible with targeted panels, where clinically actionable CNVs may span only a few hundred base pairs.

Another class of methods, including Canoes, CoNIFER, and XHMM, relies on large numbers of samples to perform denoising. These methods commonly leverage either (i) explicit baseline modeling of normal read counts (e.g., cn.MOPS) or (ii) singular value decomposition (SVD) to remove shared sources of technical variation (e.g., XHMM, CoNIFER). In a clinical setting, however, the need for rapid turnaround and the limited capacity of sequencing flow cells constrain batch sizes, making such requirements impractical. Many customers find the requirement of even 7 samples limiting.

Finally, many surveyed tools are further limited by their inability to detect both copy number gains and losses or by their failure to provide explicit quantification of copy num-

ber states. Taken together, these deficiencies underscore the need for methods specifically tailored to the high-depth, small-sample, and fine-resolution requirements of targeted sequencing panels used in clinical diagnostics. We provide a summary of these methods in Table 2.

Parameter Sensitivity

Here we evaluate the sensitivity of our model to its parameter choices. A potential criticism of the model in (4) is its dependence on predefined values of b, σ^2, and p. A principled approach would be to estimate these by maximizing the marginal likelihood,

$$\mathcal{L}(b, \sigma^2, p) = \log p(y|b, \sigma^2, p) \tag{12}$$

which requires integrating out the latent states x. However, this approach is impractical for two reasons. First, $\mathcal{L}$ is highly sample-dependent and unstable when evaluated on limited data. Second, because $\mathcal{L}$ lacks a closed-form expression, standard gradient-based optimization methods (Humpherys and Jarvis, 2020) cannot be applied. Alternatives such as Bayesian optimization (Frazier, 2018) are possible but introduce additional complexity and computational cost (Bengio et al., 2017). Furthermore, even if optimal parameters were obtained, they would likely need to be re-estimated for each sequencing batch due to variation in operators, reagent lots, and instruments, which is infeasible in clinical or commercial settings.

In this work, we set the parameter values based on heuristics. Given the interpretations of the parameters, a reasonable value for p was 0.04, given that there were 120 amplicons in the panel and we would normally expect between two and 4 jumps in a positive sample. We chose $\sigma^2 = 1$, to allow for large jumps, in synthetic data we found the value not particularly important. Finally, b was chosen to be the standard deviation of data commonly seen in customer data, 0.1.

However, we also demonstrate empirically that the model's output is robust to parameter specification across a wide range of plausible values. We conduct a sensitivity analysis by systematically varying b, σ^2, and p within clinically realistic bounds, computing the mean squared error (MSE) for each configuration on the held-out data described above. Each experiment is repeated 10 times to obtain confidence intervals, and the results are summarized in Figure 6. We find that CNV estimates are largely insensitive to p and σ^2 unless they are set to implausibly small values. Only b exhibits noticeable influence, reflecting its role in controlling the contribution of each amplicon to the CNV estimate. Nonetheless, within the empirically observed range (0.1–0.3), variation in b has minimal effect on MSE, aligning with the variance levels measured in our data (see Figure 6).

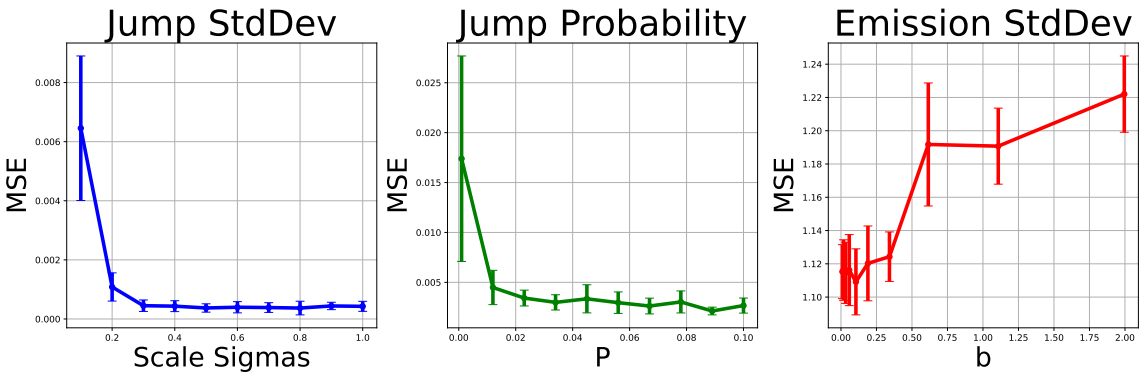

Figure 6: The left figure shows the effect that σ^2 has on the MSE of the estimated CNRs. We can see that setting this parameter to incredibly small values results in a large MSE due to missing true changes. However, beyond 0.4, the specific value is irrelevant to predictive quality. Likewise, modifying p, as shown in the middle plot, results in an inflated MSE for small values, but beyond 0.02, the specific value has a minimal impact on the MSE. Finally, the parameter b has a "sweet spot" of 0.1; values too small result in overfitting the trajectory, while too large of values result in no information being incorporated. However, around this optimal value there is a wide range where specific values have minimal impact on the quality of fit.