Combining Bayesian and Frequentist Inference for Laboratory-Specific Performance Guarantees in Copy Number Variation Detection

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

Alex V. Kotlar KOTLARA@PILLARBIOSCI.COM
Pillar Biosciences Inc
Natick, MA, USA

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

Abstract

Targeted amplicon panels are widely used in oncology diagnostics, but providing per-gene performance guarantees for copy number variant (CNV) detection remains challenging due to amplification artifacts, process-mismatch heterogeneity, and limited validation sample sizes. While Bayesian CNV callers naturally quantify per-sample uncertainty, translating this into the frequentist population-level guarantees required for clinical validation, coverage rates, false-positive bounds, and minimum detectable copy-number changes, is a fundamentally different inferential problem. We show empirically that even robust Bayesian credible intervals, including coarsened posteriors and sandwich-adjusted intervals, are severely miscalibrated on panels with small amplicon counts per gene. To address this, we propose a hybrid framework that evaluates Bayesian posterior functionals on validation samples and models the resulting squared losses with a Gamma distribution, yielding tolerance intervals with valid frequentist coverage. Three components make the method practical under real-world constraints: (1) imputation that removes the influence of true CNV-positive samples without requiring known ground truth, (2) regularization to address small sample variability, and (3) evidence-based stratification on the log model evidence to accommodate non-exchangeable noise profiles arising from process mismatch. Evaluated on two targeted amplicon panels using leave-one-out cross-validation, the proposed method achieves single-digit mean absolute coverage error across all genes under both process-matched and unmatched conditions, whereas Bayesian comparators exhibit mean absolute errors exceeding 0.60 on clinically relevant genes such as ERBB2.

1. Introduction

A false-positive ERBB2 amplification call can result in a breast cancer patient being prescribed trastuzumab (Wolff et al., 2018), a targeted therapy that carries a 4.0% rate of cardiac events at seven-year follow-up even in patients with confirmed ERBB2-positive disease (Romond et al., 2012), all for a molecular subtype she does not carry (Fehrenbacher et al., 2020). Conversely, a missed MET amplification in a non-small-cell lung cancer

patient can leave her on an EGFR-targeted regimen that has already been rendered ineffective by the undetected resistance mechanism, delaying the switch to MET-directed therapy while the tumor continues to progress. In both cases the analytical error is upstream: the sequencing-based diagnostic that informed the treatment decision was either insufficiently sensitive or insufficiently specific. The clinician had no way to know this, because the test report carried generic performance guarantees based on internal data, as opposed to "for this panel, in this run, the false-positive rate for ERBB2 copy-number calling is below $x\%$ with $y\%$ confidence." **Providing exactly that kind of individualized guarantee is the purpose of this work.**

The targeted amplicon panels this method was developed for offer considerable advantages compared to whole genome sequencing: quick turnaround times, compatibility with low DNA content (critical for later-stage patients who cannot donate much blood), tolerance of substantial sample degradation, and cost-effectiveness (Sie et al., 2014; Gao et al., 2019). The ability to produce results under these demanding conditions has proven invaluable to thousands of patients. However, this same setting creates a challenging environment for analysis, particularly for CNV detection, where limited amplicons (Boeva et al., 2014), process-matching difficulties (Derouault et al., 2020), and amplification artifacts (Peng et al., 2015) all degrade analytical performance.

BayesCNV (Talbot et al., 2026), the CNV method used in this work, is fundamentally Bayesian. Bayesian methods are a natural fit for diagnostics (Bours, 2021). They naturally incorporate prior information (Lee, 2024), quantify uncertainty (Gelman and Carpenter, 2020), and evaluate model evidence (Lotfi et al., 2022), the last of which serves as an indicator of questionable samples (Talbot and Ke, 2026). At the individual level, these qualities are irreplaceable.

However, providing per-gene performance guarantees for a given laboratory is a fundamentally frequentist question: it asks what sensitivity, false-positive rate, or estimation error a clinician can expect across future samples processed on that platform. Bayesian credible intervals can exhibit poor frequentist coverage under model misspecification — the posterior concentrates with a covariance differing from the frequentist sandwich variance (Kleijn and van der Vaart, 2012), yielding strictly higher risk (Müller, 2013), and misspecification can even render the posterior inconsistent (Grünwald and van Ommen, 2017). In the amplicon setting, such misspecification is a practical certainty, as no parametric model fully captures the target-specific biases of PCR amplification.

Fortunately, clinical validation protocols require running multiple samples in an initial run, providing genuine repeated observations from which we can evaluate any posterior functional of interest to obtain an empirical sampling distribution with valid frequentist coverage regardless of model correctness. Due to difficulties in procuring samples with known CNV status, we instead use the discrepancy between the posterior mean and the CNV-neutral value of 1 to evaluate variability. We develop a novel method to reduce the influence of true positives via conditional imputation of the largest CNVs in log space, then model the resulting losses with a gamma distribution, a technique adapted from actuarial practice. Given the limited number of samples typically available, we also implement a conjugate prior composed of pseudo-observations based on previous laboratory results. The resulting procedure is robust to a subset of samples containing true positives, compu-

tationally tractable, reliable under small sample sizes, and naturally accommodates prior information from previous experiments.

To better characterize performance when samples are not process-matched, we augment our method to stratify samples based on how well they align with model assumptions, quantified via the Bayesian evidence. Prior work has shown that the Bayesian evidence serves as an indicator of process mismatch or low sample quality (Talbot et al., 2025; Talbot and Ke, 2026). When the sample size is sufficiently large, we can split the samples into two groups based on evidence score and compute tolerance intervals using the above method on each group.

We evaluate the proposed framework using two targeted-amplicon validation datasets designed to examine complementary practical challenges. The first is an LBX dataset consisting of 32 FFPE samples analyzed using a 172-amplicon panel targeting five CNV-associated genes. Ten of the samples were prepared from substantially fragmented FFPE DNA, whereas the remaining 22 were prepared from higher-quality clinical FFPE material. By varying the composition of the normal reference set, this dataset allows us to evaluate coverage under process-matched, process-unmatched, and heterogeneous conditions. The second dataset was generated using an FFPE solid-tumor panel containing 341 amplicons, of which 240 target 14 genes used for CNV detection. This validation experiment contains 14 samples, including multiple known ERBB2-, EGFR-, and MET-positive standards, and is used to evaluate the effect of CNV-positive contamination and the performance of the proposed imputation procedure. Together, these datasets test the method under two common laboratory-validation complications: non-exchangeable noise caused by process mismatch and limited access to CNV-negative validation material.

Using leave-one-out evaluation, we show empirically that (1) Bayesian credible intervals, including intervals obtained using robust Bayesian modifications, have inadequate frequentist coverage; (2) tolerance limits estimated from validation-sample losses provide reliable coverage under both matched and noisy conditions; and (3) evidence-based stratification and conditional imputation are important when validation samples are non-exchangeable or include true CNV-positive samples, respectively.

The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 describes the proposed methodology. Section 4 evaluates the method using synthetic data, and Section 5 presents the data and Section 6 the corresponding results on the LBX and FFPE validation datasets. Section 7 provides concluding remarks and directions for future work. Python code implementing the models and reproducing the results is available at https://github.com/Pillar-Biosciences-Inc/CustomPerformanceCNV.

Generalizable Insights about Machine Learning in the Context of Healthcare

- **Robust Bayesian methods are insufficient for frequentist guarantees.** Even with robustified posteriors, hard-to-model noise sources can prevent reliable achievement of the nominal coverage rates required for clinical use.
- **Non-IID noise profiles can make error bounds inaccurate.** Methods that assume all samples have similar noise profiles can provide misleading guarantees when this assumption is violated.

- **Bayesian and frequentist methods are complementary.** Bayesian methods excel at the individual level, allowing the incorporation of prior information and evaluation of model assumptions. Frequentist methods are a natural choice for population-level performance guarantees needed for product validation.

2. Related Work

Read-depth-based CNV detection from targeted sequencing spans several methodological families. Normalization-then-segmentation pipelines, such as ONCOCNV Boeva et al. (2014), CNVkit Talevich et al. (2016), CONTRA Li et al. (2012), correct for GC content and library-size biases before applying CBS or threshold-based calling. PCA/SVD denoising methods such as XHMM Fromer et al. (2012) and CoNIFER Krumm et al. (2012) remove batch effects from the coverage matrix. Probabilistic count models jointly estimate artifacts and copy-number states: CODEX2 Jiang et al. (2018) via Poisson latent factors and GATK-gCNV Babadi et al. (2023) via a negative-binomial hierarchical HMM. Mixture-of-Poissons methods, cn.MOPS Klambauer et al. (2012) and panelcn.MOPS Povysil et al. (2017), model cross-sample variation as discrete copy-number components, while ExomeDepth Plagnol et al. (2012) and DECoN Fowler et al. (2016) use beta-binomial likelihoods with optimized reference sets. Benchmarking studies consistently report dataset- and laboratory-dependent performance Moreno-Cabrera et al. (2020); Munté et al. (2025), but none of these tools provide laboratory-customized frequentist performance guarantees.

An alternative strategy is to build robustness directly into the modeling stage. Power posteriors raise the likelihood to an exponent $\tau < 1$ to reduce sensitivity to misspecification Miller and Dunson (2019); Holmes and Walker (2017), and generalized Bayesian inference replaces the log-likelihood with an arbitrary loss, with learning rates calibrated via the SafeBayes framework Bissiri et al. (2016); Grünwald and van Ommen (2017). These approaches improve per-sample posterior inference but address a fundamentally different problem: they seek a better posterior for a single observation under a misspecified model, whereas we require frequentist guarantees, including coverage, sensitivity, mean squared error, that hold across a population of future samples.

3. Methods

3.1. Background and Statistical CNV Model

We consider N samples assayed on a targeted amplicon panel spanning J genes, where gene j is tiled by n_j amplicons ($j = 1, \ldots, J$). For a given sample let $s_{j,k}$ and $r_{j,k}$ denote the read counts at amplicon k of gene j for the test and diploid reference samples, respectively. Because the model is fit independently to each sample, we suppress the sample index throughout this subsection.

3.1.1. Features

For each amplicon we form a log copy-number ratio (lCNR) relative to the normal baseline. Amplicon-specific capture biases cancel in the ratio, isolating the copy-number signal. The raw lCNR is

$$\tilde{X}_{j,k} \;=\; \log\big(s_{j,k} + c\big) - \log\big(r_{j,k} + c\big), \tag{1}$$

where $c > 0$ is a small pseudo-count guarding against zero counts. We remove global depth differences by median-normalizing across all amplicons:

$$X_{j,k} \;=\; \tilde{X}_{j,k} - \text{median}_{j',k'}\big(\tilde{X}_{j',k'}\big). \tag{2}$$

After normalization, $X_{j,k} = 0$ indicates a copy-neutral amplicon, negative values indicate deletions, and positive values indicate amplifications. The median is robust to the sparse copy-number alterations expected in any single sample.

3.1.2. Statistical Model

We model the normalized lCNRs with the hierarchical Bayesian framework BayesCNV (Talbot et al., 2026). Gene-level means μ_j are partially pooled toward a global mean μ_0 through a between-gene variance σ^2. Amplicon-level noise is factored into a global scale τ_0^2 and gene-specific multipliers z_j^2, giving per-gene scales $\tau_j^2 = \tau_0^2 z_j^2$. The generative model is:

$$\begin{aligned}
\mu_0 &\sim \mathcal{N}\big(0,\ 10^2\big), & \sigma^2 &\sim \text{Inv-Gamma}(\alpha_\sigma,\ \beta_\sigma),\\
\mu_j \mid \mu_0, \sigma^2 &\sim \mathcal{N}\big(\mu_0,\ \sigma^2\big), & j &= 1, \ldots, J,\\
\tau_0^2 &\sim \text{Inv-Gamma}(\alpha_{\tau_0},\ \beta_{\tau_0}), & z_j^2 &\sim \text{Inv-Gamma}(\alpha_\tau,\ \beta_\tau),\\
\tau_j^2 &= \tau_0^2 z_j^2, & &\\
X_{j,k} \mid \mu_j, \tau_j &\sim \text{SoftLaplace}\big(\mu_j,\ \tau_j\big), & k &= 1, \ldots, n_j,
\end{aligned} \tag{3}$$

where SoftLaplace (Bingham et al., 2019) is a heavy-tailed location–scale family providing robustness to outlier amplicons. Throughout, $\mathcal{N}(\mu,\ \sigma^2)$ denotes a Normal distribution with mean μ and *variance* σ^2. The inverse-gamma hyperparameters are set to mildly regularizing values that keep the posterior proper while exerting minimal influence in data-rich genes.

3.1.3. Inference

The inferential target is the gene-level mean μ_j: values near zero indicate diploid copy number; substantial deviations signal deletions or amplifications. Posterior samples are drawn via Hamiltonian Monte Carlo (Betancourt, 2017) from $p(\theta \mid \mathbf{X})$, with $\theta = \big\{\mu_0, \sigma^2, \{\mu_j, z_j^2\}_{j=1}^J, \tau_0^2\big\}$. From each marginal posterior of μ_j we form $(1-\gamma)$ highest-posterior-density credible intervals and flag gene j as altered whenever the interval excludes zero. We approximate the marginal likelihood using a Laplace approximation, as opposed to the thermodynamic integration used in the original work. Further details appear in Talbot et al. (2026).

3.2. Removing Positives with Conditional Order Imputation

In a clinical deployment the caller is validated on K samples processed by the client laboratory. Ideally one would evaluate sensitivity and specificity against known positive and negative controls, but obtaining such controls at scale is rarely feasible. We therefore adopt a strategy that requires no prior knowledge of which samples carry copy-number alterations, exploiting the sparsity of CNV events: across the J genes of the panel, the vast majority of gene–sample pairs are diploid.

For any gene j, the K posterior means $\hat{\mu}_j^{(1)}, \ldots, \hat{\mu}_j^{(K)}$ are dominated by a diploid component near zero, with at most m samples harboring a genuine amplification. The top-m values—those most likely to reflect true gains—are the primary source of contamination in a tolerance limit estimated from the empirical distribution. We impute these suspect values from the bulk (CNV-neutral) distribution, constructing a pseudo-diploid reference from which calibrated tolerance limits can be derived. This correction is deliberately one-sided: copy-number gains can be arbitrarily large, inflating the upper tail, whereas due to low tumor purity deletions, even homozygous losses, depress $\hat{\mu}_j^{(s)}$ only modestly.

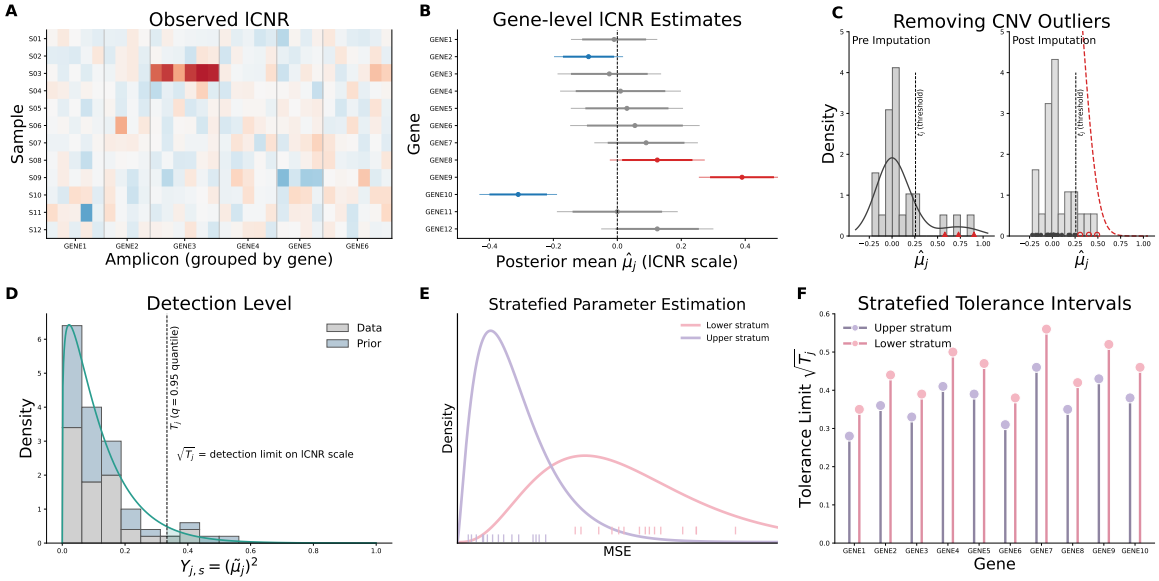

Figure 1: **(A)** Heatmap of normalized lCNRs $X_{j,k}$ for representative samples (rows) across amplicons grouped by gene (columns). **(B)** Forest plot of BayesCNV posterior means $\hat{\mu}_j$ with $(1-\gamma)$ HPD credible intervals for one representative sample. **(C)** Conditional order-statistic imputation for one representative gene. *Left:* distribution of K posterior means with top-m suspect values (filled red markers) and threshold t_j (dashed line). *Right:* distribution after replacing suspect values with order-statistic draws. **(D)** Gamma model fit (solid colored curve) to the empirical distribution of squared imputed posterior means $Y_{j,s}$ (grey) and pseudo observations (pink); the vertical dashed line marks the tolerance quantile T_j, and $\sqrt{T_j}$ is the minimum detectable lCNR deviation. **(E)** Stratified parameter estimates each sample group divided by model evidence. **(F)** Per-gene tolerance limits $\sqrt{T_j}$ in upper and lower evidence strata.

Fix gene j and write the K ordered posterior means as

$$\hat{\mu}_{j,(1)} \leq \hat{\mu}_{j,(2)} \leq \cdots \leq \hat{\mu}_{j,(K)}. \tag{4}$$

Fix a small integer $m < K$, set $L = K - m$, and define the threshold

$$t_j \;=\; \hat{\mu}_{j,(L)}, \tag{5}$$

the largest value among the L retained observations. The top-m values $\hat{\mu}_{j,(L+1)}, \ldots, \hat{\mu}_{j,(K)}$ are replaced by the following procedure, applied independently per gene.

1. **Bulk model fit.** Fit a Normal distribution $\hat{F}_j = \mathcal{N}\big(\hat{\xi}_j,\, \hat{\omega}_j^2\big)$ to the L retained values using robust estimators:

$$\hat{\xi}_j = \text{median}\big(\hat{\mu}_{j,(1)}, \ldots, \hat{\mu}_{j,(L)}\big), \qquad \hat{\omega}_j = 1.4826\ \text{MAD}\big(\hat{\mu}_{j,(1)}, \ldots, \hat{\mu}_{j,(L)}\big), \tag{6}$$

   where MAD is the median absolute deviation and $1.4826 = 1/\Phi^{-1}(3/4)$ is the Fisher consistency factor for Gaussian scale estimation (see Appendix A; Huber and Ronchetti, 2009).

2. **Truncated sampling.** Let $S_j = 1 - \hat{F}_j(t_j)$. Draw $U_1, \ldots, U_m \overset{\text{i.i.d.}}{\sim} \text{Uniform}(0, 1)$ and set

$$Z_i = \hat{F}_j^{-1}\Big(\hat{F}_j(t_j) + U_i\, S_j\Big), \qquad i = 1, \ldots, m. \tag{7}$$

   Each Z_i is a draw from $\hat{F}_j$ truncated to $[t_j, \infty)$.

3. **Order-statistic replacement.** Sort the draws to obtain $Z_{(1)} \le \cdots \le Z_{(m)}$ and set

$$\tilde{\mu}_{j,(L+i)} = Z_{(i)},\ i = 1, \ldots, m; \qquad \tilde{\mu}_{j,(r)} = \hat{\mu}_{j,(r)},\ r = 1, \ldots, L. \tag{8}$$

This procedure reproduces the classical conditional distribution of upper order statistics given $X_{(L)} = t$ for an i.i.d. continuous sample (David and Nagaraja, 2004), with the true F replaced by the robust estimate $\hat{F}_j$.

3.3. Tolerance Intervals Via a Gamma Quantile

Having removed the influence of up to m positive samples using the method in Section 3.2, we now work to obtain tolerance intervals on the posterior means by gene. Note, to maintain uncertainty in the imputations, this procedure will be repeated multiple times with different imputed values and use the empirical mean estimate. However, for clarity, we will suppress this in the following sections.

We characterize the CNV performance via the MSEs of the posterior means (assuming the sample is CNV neutral). That is

$$Y_{j,s} \;=\; \big(\tilde{\mu}_j^{(s)}\big)^2. \tag{9}$$

A natural model for Y is a Gamma distribution. If $\tilde{\mu}_j^{(s)} \sim \mathcal{N}(0,\, \tau_j^2)$ then $Y_{j,s} \sim \tau_j^2\, \chi_1^2 = \text{Gamma}\big(\frac{1}{2},\, 2\tau_j^2\big)$. More generally, a residual process-mismatch bias δ_j gives

$$\tilde{\mu}_j^{(s)} \sim \mathcal{N}\big(\delta_j,\, \tau_j^2\big) \quad \Longrightarrow \quad Y_{j,s} = \tau_j^2\, \chi_1^2(\lambda_j), \qquad \lambda_j = \delta_j^2/\tau_j^2, \tag{10}$$

a scaled noncentral chi-squared variate (Urkowitz, 1967). Moment-matching to Gamma(α_j, s_j) yields (see Appendix A for the derivation)

$$\alpha_j = \frac{(1+\lambda_j)^2}{2(1+2\lambda_j)}, \qquad s_j = \frac{2\tau_j^2(1+2\lambda_j)}{1+\lambda_j}. \tag{11}$$

When $\delta_j = 0$ we recover $\alpha_j = \frac{1}{2}$; a nonzero bias inflates α_j in proportion to λ_j, so a flexible Gamma MLE accommodates both regimes. Additional motivation for the Gamma family is given in Appendix A.

Once we estimate the parameters of the gamma distribution, we construct a $(1-p)$ upper tolerance bound as the corresponding quantile of the fitted distribution,

$$T_j = F^{-1}_{\text{Gamma}(\hat{\alpha}_j, \hat{s}_j)}(1-p). \tag{12}$$

Because the squared-error loss $Y_{j,s} = \hat{\mu}_j^2$ is related to the original posterior mean by a monotonic transformation, the tolerance bound on the loss scale translates directly to one on the CNV estimate: a new diploid sample will satisfy $|\hat{\mu}_j^{\text{new}}| \leq \sqrt{T_j}$ with probability at least $1-p$, i.e.,

$$\Pr\left(\hat{\mu}_j^{\text{new}} \notin \left[-\sqrt{T_j}, \sqrt{T_j}\right]\right) < p. \tag{13}$$

This bound is two-sided and symmetric by construction; however, when a nonzero process-mismatch bias $\delta_j \neq 0$ is present, the true exceedance probability will be unequally distributed between the upper and lower tails, even though the aggregate coverage guarantee remains valid.

While this is framed as a bound on the expected variability of CNV estimates, we also can use this as a proxy for expected performance for positive samples as well. Provided that the amplicon-level biases and variability is independent of CNV status, this is roughly the certainty we would see in lCNR estimates of positive samples as well. In particular, $\exp(\sqrt{T_j})$ is the lowest detectable CNV for gene j.

3.4. Incorporating a Conjugate Prior with Pseudo-Observations

Quantile estimation can be inherently noisy, particularly in the upper quantiles like those used in this work. This is a particularly acute issue in commercial applications, as customers often repeatedly use different combinations of samples, both to evaluate robustness and also search for combinations that yield the "best" results. This can result in a particularly difficult combination when this is run at low sample sizes ($N = 10$). Our applications require low variance, even at the expense of a higher MSE due to increased bias.

We achieve this by placing a prior on both α_j and s_j. We can create a conjugate prior for any regular model where $A = \{x : p(\theta|x) > 0\}$ does not depend on θ, that is support does not depend on the parameters, by the incorporation of pseudo observations $(\xi_1, \ldots, \xi_{K'})$ (Bickel and Doksum, 2015). That is, our prior is $p(\alpha_j, s_j) \propto \prod_{k=1}^{K'} p(\xi_k|\alpha_j, s_j)$ resulting in a posterior of

$$p(\alpha_j, s_j|Y_{j,s}, \{\xi\}) = \prod_{k=1}^{K'+K} p(\xi'_k|\alpha_j, s_j) \Big/ \int \prod_{k=1}^{K'+K} p(\xi'_k|\alpha_j, s_j), \tag{14}$$

where $\{\xi'_k\}_{k=1:K+K'} = \{Y_{j,1} \ldots, Y_{j,s}, \xi_1, \ldots, \xi_{K'}\}$. The advantage of this posterior is that computation is straightforward with MAP estimation, simply augment the observed data with pseudo-observations and perform maximum likelihood.

While the easiest method for the above prior is to simply tack on a small number of samples, a less noisy method is to incorporate a large number of samples and weight their contribution in the likelihood to achieve a fixed sample size. We propose two methods for choosing these prior samples. The first is to use samples from a previous related experiment. This method has the downside that it either (i) requires the customer to run a first experiment without a prior or (ii) involves using one customer's data for a separate group. The second is to use an internal run to estimate α_j and s_j under standard conditions, then generate a large number of pseudo observations with those parameters. This is the method we have chosen commercially.

3.5. Evidence-Stratified Tolerance Limits

The tolerance limits in Section 3.3 assume that the K samples are exchangeable draws from a single production population. We frequently find that not all samples are process matched (i.e. sequenced with the same instrument, library preparation, and analysis pipeline used clinically), introducing bias and excess variance (Appendix B). When samples are pooled, the squared posterior means $Y_{j,s}$ follow a mixture of Gamma-like distributions whose components may differ in both location and scale due to different noise profiles. Fitting a single Gamma(α_j, s_j) to such a mixture can yield tolerance quantiles that differ dramatically from the true value of each group individually.

We eliminate this artifact by stratifying the validation samples on the log model evidence $Z_s = \log p(\mathbf{X}_s \mid \mathcal{M})$, which serves as a process-matching diagnostic (Talbot and Ke, 2026): well-matched samples yield high evidence while mismatched samples yield low evidence. There are a variety of techniques for approximating this value, such as thermodynamic integration (Gelman and Meng, 1998) or annealed importance sampling (Neal, 2001), or in this work a Laplace approximation (Gelman et al., 1995). Partitioning on $\mathcal{Z}_s$ produces strata within which the single-Gamma model is approximately correctly specified.

Concretely, let Z_{med} denote the sample median of $\{Z_1, \ldots, Z_K\}$. We define two strata:

$$\mathcal{S}^+ = \big\{s : Z_s \geq Z_{\text{med}}\big\}, \qquad \mathcal{S}^- = \big\{s : Z_s < Z_{\text{med}}\big\}, \tag{15}$$

with $|\mathcal{S}^+| = \lceil K/2 \rceil$ and $|\mathcal{S}^-| = \lfloor K/2 \rfloor$. Within each stratum $g \in \{+, -\}$ and for each gene j, we fit a separate Gamma model to the squared imputed posterior means $\{Y_{j,s} : s \in \mathcal{S}^g\}$, augmented with the pseudo-observations described in Section 3.4, yielding per-stratum tolerance quantiles

$$T_j^{(g)} = F^{-1}_{\text{Gamma}(\hat{\alpha}_j^{(g)}, \hat{s}_j^{(g)})}\big(1 - p\big), \qquad g \in \{+, -\}. \tag{16}$$

The reported tolerance limit for gene j in each sample would then be $T_{j,s} = T_j^{g(s)}$.

This procedure requires $K > 20$ so that each stratum retains at least ten observations, which in our experience is the number required to accurately estimate the parameters of the gamma distribution. This stratification is also optional, one potential method is to detect whether there is evidence for non-exchangability in the samples via a test like the

one developed in Talbot and Ke (2026), allowing for formal assessment of whether such stratification is even necessary.

4. Synthetic Results

The purpose of this synthetic section is to evaluate the impact that the imputation scheme and prior have on the quantile estimates. We evaluate the bias, variance, and mean squared error method as a function of sample size, with a particular focus on the contributions of the prior distribution and the outlier-imputation scheme. Distribution parameters $\delta = 0.2$ and $\tau^2 = 0.17$ were estimated from a held-out panel analyzed in Section 6.2; importantly, these values are drawn from a different panel than the one used to construct the informative prior, ensuring that the prior is not fit to the evaluation data. Synthetic datasets were repeatedly drawn at sample sizes ranging from $N = 5$ to $N = 50$ from $\mathcal{N}(\delta, \tau^2)$, and for each draw we computed the true quantile bounding $|X|$, comparing four estimators. First the standard maximum-likelihood estimate (Empirical), used as a baseline. Second, our novel estimator with no prior information and $m = 1$. Third, the same method but with the informative prior and an effective sample size of 5. Finally, our estimator with the same informative prior and an imputation weight that scales linearly with N, representing the regime in which the number of outliers grows proportionally with the number of samples processed. The true value was 0.665.

Results are displayed in Figure 2. Relative to the empirical baseline, the proposed estimator without a prior reduces bias and MSE at the cost of increased variance; at small sample sizes the standard error reaches approximately 25% of the true parameter value, a level that is prohibitively high for reliable clinical decision-making. Incorporating the informative prior yields a substantial reduction in variance, roughly 2.5×, bringing the relative standard error to approximately 10%, a range consistent with stable performance guarantees under the variability introduced by routine sample-selection interventions. Finally, the $m = 0.2N$ condition demonstrates that the imputation scheme remains well-calibrated even when the outlier proportion grows with sample size: neither bias nor MSE increases materially relative to the fixed-m setting, confirming that the approach is robust to elevated rates of true-positive contamination.

5. Cohorts

We evaluate our method across two targeted amplicon panels, spanning a range of amplicon counts, sample sizes, and degrees of process mismatch (Table 1).

Small Panel LBX. This dataset comprises 32 FFPE samples sequenced with a 172-amplicon panel designed for low-cost cancer screening, targeting five CNV-associated genes. Of these samples, 10 libraries were prepared from substantially fragmented FFPE DNA, and 22 from higher-quality clinical FFPE material. To generate matched and unmatched evaluation conditions, we chose 5 of the 32 samples to construct an averaged normal reference, with the fraction of fragmented DNA in the reference ranging from 0 to all 5 samples; the caller was then evaluated on the remaining high-quality samples. This construction was repeated 30 times with different normal selections, allowing us to carefully control the

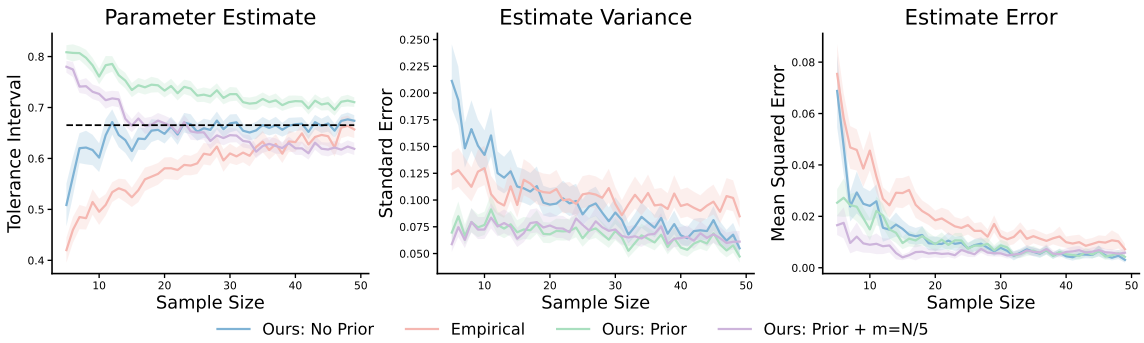

Figure 2: Interval estimation performance as a function of sample size for the four estimators. **Left:** posterior mean parameter estimate with 95% confidence intervals. **Center:** standard error of each estimator. **Right:** mean squared error. Data were generated from $\mathcal{N}(\delta, \tau^2)$ with $\delta = 0.2$ and $\tau^2 = 0.17$ over $N \in [5, 50]$.

level of known process mismatch. For each sample we additionally evaluated the marginal likelihood using a Laplace approximation.

This same cohort also supports a heterogeneous-mixture analysis (Section 6.1.3), in which we pool the matched and unmatched conditions above into a single run of 39 samples (17 process matched, 22 not) to evaluate the benefit of evidence-based stratification when process mismatch is not known in advance. This is a re-partitioning of the same Small Panel LBX samples across trials rather than an independent dataset.

FFPE Validation Data. Our second cohort is drawn from a 341-amplicon panel designed for FFPE solid tumor diagnostics, of which 240 amplicons target 14 CNV-relevant genes. This internal validation run comprised 14 samples: 5 NIST standard samples CNV-positive for ERBB2, and 5 positive for EGFR and MET. Because samples with known CNVs are difficult to procure, validation runs of this kind necessarily enrich for positives, with a correspondingly high fraction of samples carrying CNVs in the same gene, a condition that can distort the baseline and obscure true signal, motivating the imputation analysis in Section 6.2.

Table 1: Summary of evaluation cohorts.

Cohort	Panel size	Genes	Samples	Composition
Small Panel LBX	172 amplicons	5	32	10 fragmented / 22 high-quality
FFPE Validation	341 amplicons	14	14	5 ERBB2+ / 5 EGFR & MET+ NIST

6. Results

We evaluate the empirical coverage of the proposed tolerance intervals against the standard Bayesian credible interval and two alternatives drawn from the Bayesian robust-

ness literature, along with a traditional MSE-based coverage estimate corresponding to a Gamma$(1/2, s)$ distribution. All methods are assessed on a common criterion: for each gene j and each diploid validation sample s, we record whether the interval excludes zero, and report the resulting false-positive rate over (j, s) pairs relative to the nominal level γ. The first comparator is the standard $(1-\gamma)$ highest-posterior-density (HPD) credible interval for μ_j^{new} obtained directly from BayesCNV. The remaining two comparators apply misspecification-robust modifications to the BayesCNV posterior for each new sample and are the coarsened posterior of Miller and Dunson (2019), and the sandwich posterior of Müller (2013).

We evaluate coverage using leave-one-out cross-validation: for each validation sample s, the tolerance interval is estimated from the remaining $K-1$ individuals and we record whether $|\hat{\mu}_j^{(s)}| \leq \sqrt{T_j}$. Empirical coverage is then reported alongside interval width for each method, with width serving as the secondary criterion: narrower intervals that maintain nominal coverage correspond directly to a lower minimum detectable CNV $\exp(\sqrt{T_j})$ and hence greater clinical sensitivity.

6.1. Small Panel LBX

6.1.1. Process Matched Coverage

We quantify calibration via the *mean absolute coverage error* (MACE),

$$\text{MACE} = \frac{1}{|\Gamma|} \sum_{\gamma \in \Gamma} \left| \hat{C}(\gamma) - \gamma \right|, \tag{17}$$

where Γ is a uniform grid of nominal coverage levels on $[0.7, 1]$ and $\hat{C}(\gamma)$ is the empirical coverage at level γ evaluated per gene over leave-one-out folds; values are reported multiplied by 100 (Table 2).

Under matched conditions all three BayesCNV-based comparators are severely miscalibrated. The standard HPD and sandwich (Müller, 2013) intervals are drastically overconfident, with their MACE values reaching up to 66 and 69 respectively; these reflect intervals so narrow that they fail to contain the true signal across most of the $[0.7, 1]$ coverage range, the expected consequence of applying a large-sample posterior approximation at the small amplicon counts ($n_j \approx 4$–20) of this panel. The coarsened posterior (Miller and Dunson, 2019) is erratic: near-zero MACE on some genes indicates the opposite failure mode, intervals so wide that empirical coverage is 100% throughout, while MACE as high as 42 on other genes indicates severe overconfidence, with both failure modes present simultaneously depending on the gene.

The proposed tolerance intervals and the MSE-based estimator together form a clearly distinct tier, each achieving single-digit MACE across all five genes and representing a substantial improvement over all three Bayesian comparators. Within this tier, the proposed intervals are the best or tied-best on every gene, with the largest margin on ERBB2 (3.0 ± 0.2 vs. 5.6 ± 0.5 for MSE), a gene of direct clinical relevance for copy-number calling.

6.1.2. Unmatched Coverage

Under maximum process mismatch the failure modes of the Bayesian comparators become more varied and in some respects more severe, as shown in Table 2 and Figure 3. Process

mismatch inflates the posterior variance on several genes, causing the standard, coarsened, and sandwich intervals to swing from overconfidence to near-total underconfidence (empirical coverage approaching 100%, MACE near zero) on genes such as KIT and PIK3CA, a change in failure direction, not a genuine improvement in calibration. On ERBB2, where mismatch does not rescue the posterior variance, all three Bayesian variants remain severely overconfident (MACE $\times$ 100 up to 74). This instability, intervals that are either far too wide or far too narrow depending on the gene and the degree of process mismatch, renders the Bayesian comparators unreliable for clinical use.

The proposed tolerance intervals remain uniformly well-calibrated across all five genes, as does the MSE-based estimator, which degrades modestly but remains in the single-digit MACE regime. The proposed method is again the best or near-best on every gene, demonstrating that robustness to process mismatch is achieved by construction rather than by fortuitous cancellation of errors.

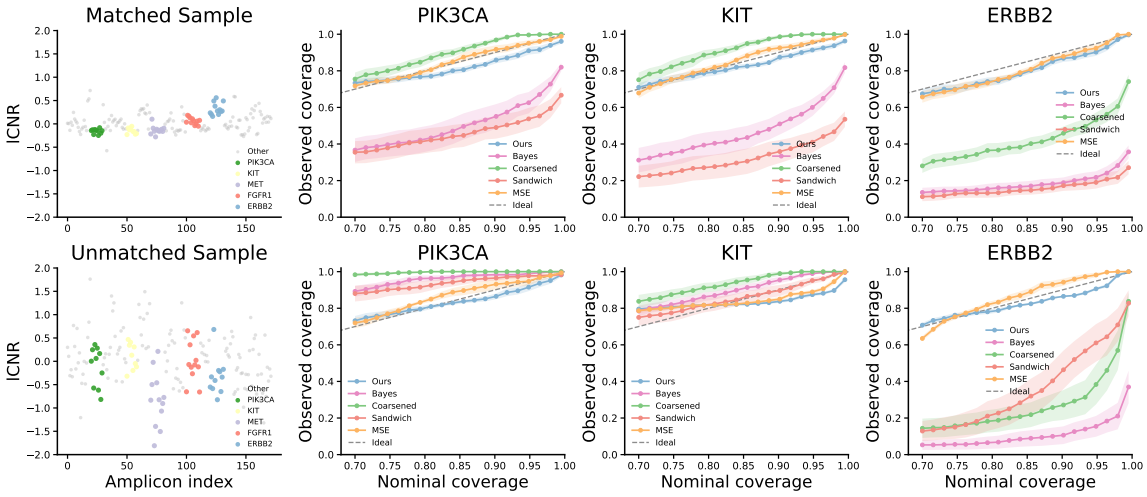

Figure 3: Calibration curves of the various methods. **Top:** Coverages in the process-matched experiments. The far left plots the lCNR of a representative process-matched sample, with amplicons targeting the five CNV-relevant genes shown with different colors. The remaining plots show the true coverages on three of the five genes as a function of target coverages. The ideal is shown via a dotted line, while the average coverage and 95% confidence interval for each of the five considered methods is shown in a different color. **Bottom:** These plots have identical interpretation, except when the data come from unmatched (highly noisy) samples.

6.1.3. Heterogeneous Mixture

We then evaluate a situation that we frequently encounter, when both matched and unmatched samples are included in the same run. This situation is evident when we examine

Table 2: Mean absolute error $\times 100$ with 95% confidence intervals for coverage evaluation on the Small Panel LBX using matched and unmatched normals. Results are shown for five methods: Gamma (Ours), MSE, Bayesian, Sandwich, and Coarsened.

	Matched					**Unmatched**				
Method	PIK3CA	KIT	MET	FGFR1	ERBB2	PIK3CA	KIT	MET	FGFR1	ERBB2
Bayes	33 ± 6	37 ± 5	40 ± 7	54 ± 2	66 ± 3	11 ± 4	12 ± 1	45 ± 7	25 ± 6	74 ± 4
Coarsened	$6.6 \pm .9$	8.0 ± 1	18 ± 5	14 ± 3	42 ± 3	15 ± 2	9.5 ± 2	45 ± 13	10 ± 2	57 ± 7
Sandwich	38 ± 6	52 ± 6	47 ± 7	60 ± 3	69 ± 3	10 ± 2	7.6 ± 1.4	43 ± 13	20 ± 4	49 ± 8
MSE	$4.2 \pm .5$	$4.2 \pm .6$	$6.4 \pm .6$	$4.2 \pm .7$	$5.6 \pm .5$	$5.7 \pm .5$	$4.4 \pm .5$	8.1 ± 2	$7.4 \pm .5$	$4.2 \pm .7$
Gamma	$4.9 \pm .2$	$4.4 \pm .3$	$6.1 \pm .6$	$5.3 \pm .4$	$3.0 \pm .2$	$4.4 \pm .6$	$5.4 \pm .1$	$5.5 \pm .7$	$7.0 \pm .8$	$3.2 \pm .2$

the distribution of likelihoods, shown in panel (A) of Figure 4. Unsurprisingly, the unmatched samples appear noisier and have correspondingly lower likelihoods.

We compared the results of our method with and without stratification. Note that since 22 samples are from one group and 17 in the other, this analysis does not "cheat" by having the median perfectly align with group membership. We first show the performance guarantees for MET in panel (B) of Figure 4. Note that the estimate does not lie between the estimates of the two groups over the entire interval; at the highest quantiles the estimate is excessively loose and larger than each group individually. This is the price we pay for modeling a mixture using a single parametric model. While we know that the quantile for the mixture is less than the maximum of the individual quantiles, with misspecified parametric models this is no longer true. We can see in panel (C) of Figure 4 that the stratified model is better over the entire range, and importantly is similar to the aggregate guarantee, while having narrower intervals. Ironically, by modeling heterogeneity in the noise distribution, we are able to accurately state that our performance is better than our estimate when such heterogeneity is ignored.

6.2. FFPE Validation Data

The effect of these outliers on the distribution of the MET losses is shown in panel (A) of Figure 5. Running this high a fraction of samples with CNVs in the same gene is inadvisable, as it can affect the baseline and obscure true signal. However, this is a common occurrence because samples with known CNVs are hard to procure, and validation runs naturally use multiple positive samples to ensure our methods work. As a result we can use this dataset to show the importance of the imputation step in Section 3.2, and evaluate the sensitivity to the imputation fraction.

To do so, we first evaluated the 95% quantile on the samples, excluding the samples with known positives on a per-gene basis, which we treat as the ground-truth value. We then fit our method on the full dataset with the positives, increasing the fraction of imputed values up to 50% of the samples. The difference between the estimated and true value is plotted in panel (B) of Figure 5.

We found that our method was relatively robust to imputation fraction, with the non-CNV genes having an average relative difference of 8% when 30% of the values are imputed,

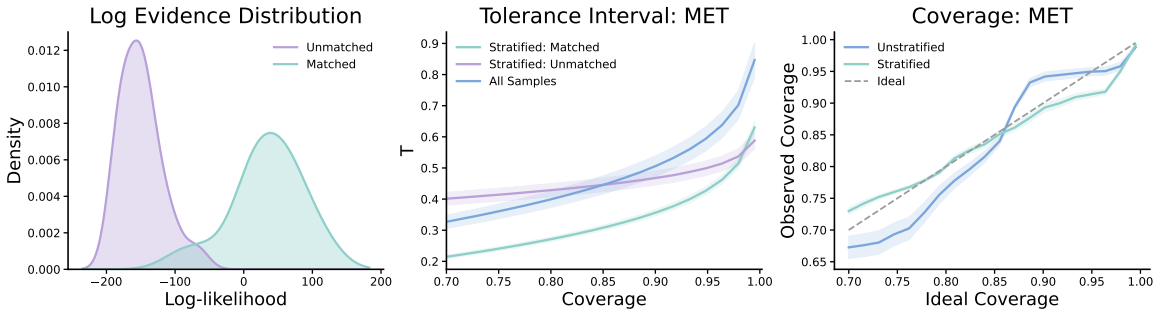

Figure 4: The effect of stratification in heterogeneous mixtures. **(A)** The distributions of log-likelihoods in the process matched and unmatched populations. **(B)** The tolerance intervals for MET on each of the subgroups for a stratified fit, and all samples when group differences are ignored. **(C)** The associated coverages for each method.

with a maximum of 22%. Even when 50% of the values are imputed the average relative difference was 15% with a maximum of 38%. While this is naturally suboptimal compared to no imputation, it is dramatically preferable to ignoring the outlier effect of true positives. If this imputation step were not performed, it would erroneously suggest that the limit of detection for MET was 8.6, as opposed to the true and far more reasonable value of 1.5 as shown in panel (C) of Figure 5.

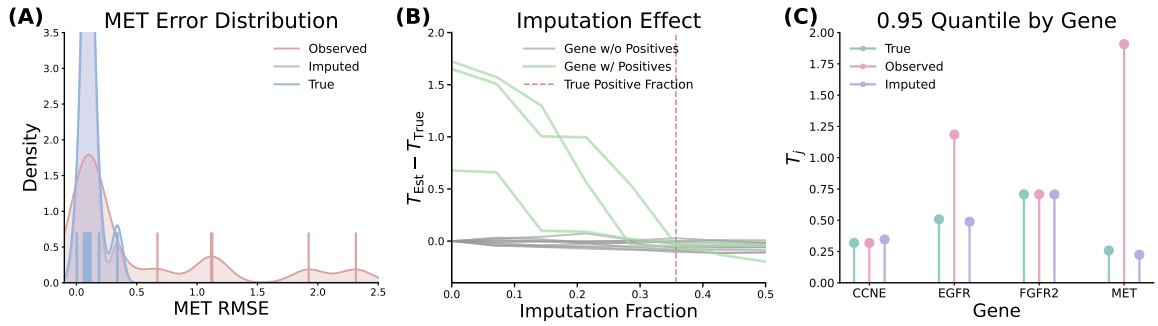

Figure 5: The effect of imputation on MET tolerance interval estimates. **(A)** plots the KDEs for true, observed, and imputed MET RMSE. **(B)** plots the difference between estimated and true 0.95 quantiles across imputation fractions, with the dashed line indicating the true positive fraction. **(C)** plots the true, observed, and imputed 0.95 quantiles for CCNE, EGFR, FGFR2, and MET.

7. Discussion

This work demonstrates that providing laboratory-specific frequentist performance guarantees for a Bayesian diagnostic pipeline requires a fundamentally different inferential strategy than improving the posterior itself. Our experiments demonstrate that state-of-the-art robust Bayesian modifications, the coarsened posterior (Miller and Dunson, 2019) and the sandwich adjustment (Müller, 2013) fail to achieve reliable frequentist coverage on the small-amplicon panels typical of clinical NGS, because calibration is a property of the repeated-sampling procedure, not of any individual posterior. The proposed framework addresses this gap by recasting the problem as tolerance-interval estimation on empirical loss distributions, incorporating regularization from previous laboratories to improve performance and reduce variance. That the simpler MSE-based estimator also enters the single-digit MACE regime reinforces the broader lesson: the critical design choice is the frequentist reframing itself, with the Gamma model providing a further edge through its theoretical grounding in the noncentral chi-squared structure of the losses and its accommodation of process-mismatch bias via the shape parameter (Appendix A). Beyond CNV detection, this hybrid strategy, Bayesian inference for individual-level uncertainty paired with frequentist calibration for population-level guarantees, is applicable to any ML-based clinical assay that requires quantifiable performance bounds. The evidence-stratification component further highlights a principle with wider relevance in healthcare ML: subgroup-specific performance reporting, much like disaggregated evaluation in algorithmic fairness, can expose heterogeneity that aggregate metrics conceal.

Several limitations point to natural avenues for future work. First, the evidence-based stratification relies on a binary median split of the log model evidence Z_s, a deliberately simple scheme that treats process mismatch as a dichotomy and requires $K > 20$ to provide adequate per-stratum sample sizes. This is perhaps the most accessible opportunity for improvement: replacing the hard partition with continuous sample weights or data-driven clustering (Talbot and Ke, 2026) would allow finer-grained accommodation of heterogeneity without halving the effective sample size per stratum. Second, the imputation method currently functions only when $m < 0.3N$, which results in underconfidence when a large proportion of samples share a CNV in a specific gene. Further work to improve the breakdown point would be useful to make this method more robust to these situations.

References

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.

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

Peter J Bickel and Kjell A Doksum. *Mathematical statistics: basic ideas and selected topics, volumes I-II package*. Chapman and Hall/CRC, 2015.

Eli Bingham, Jonathan P Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D Goodman. Pyro: Deep universal probabilistic programming. *Journal of machine learning research*, 20(28):1–6, 2019.

Pier Giovanni Bissiri, Chris C Holmes, and Stephen G Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):1103–1130, 2016.

Valentina Boeva, Tatiana Popova, Maxime Lienard, Sebastien Toffoli, Maud Kamal, Christophe Le Tourneau, David Gentien, Nicolas Servant, Pierre Gestraud, Thomas Rio Frio, et al. Multi-factor data normalization enables the detection of copy number aberrations in amplicon sequencing data. *Bioinformatics*, 30(24):3443–3450, 2014.

Martijn JL Bours. Bayes' rule in diagnosis. *Journal of Clinical Epidemiology*, 131:158–160, 2021.

Herbert A David and Haikady N Nagaraja. *Order statistics*. John Wiley & Sons, 2004.

Paco Derouault, Jasmine Chauzeix, David Rizzo, Federica Miressi, Corinne Magdelaine, Sylvie Bourthoumieu, Karine Durand, Helene Dzugan, Jean Feuillard, Franck Sturtz, et al. Covcopcan: An efficient tool to detect copy number variation from amplicon sequencing data in inherited diseases and cancer. *PLoS Computational Biology*, 16(2): e1007503, 2020.

Louis Fehrenbacher, Reena S Cecchini, Charles E Geyer Jr, Priya Rastogi, Joseph P Costantino, James N Atkins, John P Crown, Jonathan Polikoff, Jean-Francois Boileau, Louise Provencher, et al. Nsabp b-47/nrg oncology phase iii randomized trial comparing adjuvant chemotherapy with or without trastuzumab in high-risk invasive breast cancer negative for her2 by fish and with ihc 1+ or 2+. *Journal of clinical oncology*, 38(5): 444–453, 2020.

Anna Fowler, Shazia Mahamdallie, Elise Ruark, Sheila Seal, Emma Ramsay, Matthew Clarke, Imran Uddin, Harriet Wylie, Ann Strydom, Gerton Lunter, et al. Accurate clinical detection of exon copy number variants in a targeted ngs panel using decon. *Wellcome open research*, 1:20, 2016.

Menachem Fromer, Jennifer L Moran, Kimberly Chambert, Eric Banks, Sarah E Bergen, Douglas M Ruderfer, Robert E Handsaker, Steven A McCarroll, Michael C O'Donovan, Michael J Owen, et al. Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *The American Journal of Human Genetics*, 91(4): 597–607, 2012.

Meiling Gao, Maurizio Callari, Emma Beddowes, Stephen-John Sammut, Marta Grzelak, Heather Biggs, Linda Jones, Abdelhamid Boumertit, Sabine C Linn, Javier Cortes, et al. Next generation-targeted amplicon sequencing (ng-tas): an optimised protocol and computational pipeline for cost-effective profiling of circulating tumour dna. *Genome medicine*, 11(1):1, 2019.

Andrew Gelman and Bob Carpenter. Bayesian analysis of tests with unknown specificity and sensitivity. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 69 (5):1269–1283, 2020.

Andrew Gelman and Xiao-Li Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical science*, 13(2):163–185, 1998.

Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 1995.

Peter Grünwald and Thijs van Ommen. Inconsistency of bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12(4):1069–1103, 2017.

Chris C Holmes and Stephen G Walker. Assigning a value to a power likelihood in a general bayesian model. *Biometrika*, 104(2):497–503, 2017.

Peter J. Huber and Elvezio M. Ronchetti. *Robust Statistics*. Wiley, Hoboken, NJ, 2nd edition, 2009. ISBN 9780470129906.

Yuchao Jiang, Rujin Wang, Eugene Urrutia, Ioannis N Anastopoulos, Katherine L Nathanson, and Nancy R Zhang. Codex2: full-spectrum copy number variation detection by high-throughput dna sequencing. *Genome biology*, 19(1):202, 2018.

Günter Klambauer, Karin Schwarzbauer, Andreas Mayr, Djork-Arne Clevert, Andreas Mitterecker, Ulrich Bodenhofer, and Sepp Hochreiter. cn. mops: mixture of poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic acids research*, 40(9):e69–e69, 2012.

BJK Kleijn and AW van der Vaart. The bernstein-von-mises theorem under misspecification. *Electronic Journal of Statistics*, 6:354–381, 2012.

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

Se Yoon Lee. Using bayesian statistics in confirmatory clinical trials in the regulatory setting: a tutorial review. *BMC Medical Research Methodology*, 24(1):110, 2024.

Jason Li, Richard Lupat, Kaushalya C Amarasinghe, Ella R Thompson, Maria A Doyle, Georgina L Ryland, Richard W Tothill, Saman K Halgamuge, Ian G Campbell, and Kylie L Gorringe. Contra: copy number analysis for targeted resequencing. *Bioinformatics*, 28(10):1307–1313, 2012.

Sanae Lotfi, Pavel Izmailov, Gregory Benton, Micah Goldblum, and Andrew Gordon Wilson. Bayesian model selection, the marginal likelihood, and generalization. In *International Conference on Machine Learning*, pages 14223–14247. PMLR, 2022.

Jeffrey W Miller and David B Dunson. Robust bayesian inference via coarsening. *Journal of the American Statistical Association*, 114:1113–1125, 2019.

José Marcos Moreno-Cabrera, Jesús Del Valle, Elisabeth Castellanos, Lidia Feliubadaló, Marta Pineda, Joan Brunet, Eduard Serra, Gabriel Capellà, Conxi Lázaro, and Bernat Gel. Evaluation of cnv detection tools for ngs panel data in genetic diagnostics. *European Journal of Human Genetics*, 28(12):1645–1655, 2020.

Ulrich K Müller. Risk of bayesian inference in misspecified models, and the sandwich covariance matrix. *Econometrica*, 81(5):1805–1849, 2013.

Elisabet Munté, Carla Roca, Jesús Del Valle, Lidia Feliubadaló, Marta Pineda, Bernat Gel, Elisabeth Castellanos, Barbara Rivera, David Cordero, Víctor Moreno, et al. Detection of germline cnvs from gene panel data: benchmarking the state of the art. *Briefings in Bioinformatics*, 26(1):bbae645, 2025.

Radford M Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001.

Quan Peng, Ravi Vijaya Satya, Marcus Lewis, Pranay Randad, and Yexun Wang. Reducing amplification artifacts in high multiplex amplicon sequencing by using molecular barcodes. *BMC genomics*, 16(1):589, 2015.

Vincent Plagnol, James Curtis, Michael Epstein, Kin Y Mok, Emma Stebbings, Sofia Grigoriadou, Nicholas W Wood, Sophie Hambleton, Siobhan O Burns, Adrian J Thrasher, et al. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics*, 28(21):2747–2754, 2012.

Gundula Povysil, Antigoni Tzika, Julia Vogt, Verena Haunschmid, Ludwine Messiaen, Johannes Zschocke, Günter Klambauer, Sepp Hochreiter, and Katharina Wimmer. panelcn. mops: Copy-number detection in targeted ngs panel data for clinical diagnostics. *Human mutation*, 38(7):889–897, 2017.

Edward H Romond, Jong-Hyeon Jeong, Priya Rastogi, Sandra M Swain, Charles E Geyer Jr, Michael S Ewer, Vikas Rathi, Louis Fehrenbacher, Adam Brufsky, Catherine A Azar, et al. Seven-year follow-up assessment of cardiac function in nsabp b-31, a randomized trial comparing doxorubicin and cyclophosphamide followed by paclitaxel (acp) with acp plus trastuzumab as adjuvant therapy for patients with node-positive, human epidermal growth factor receptor 2–positive breast cancer. *Journal of Clinical Oncology*, 30(31): 3792–3799, 2012.

Daoud Sie, Peter JF Snijders, Gerrit A Meijer, Marije W Doeleman, Marinda IH van Moorsel, Hendrik F van Essen, Paul P Eijk, Katrien Grünberg, Nicole CT van Grieken, Erik Thunnissen, et al. Performance of amplicon-based next generation dna sequencing for diagnostic gene mutation profiling in oncopathology. *Cellular Oncology*, 37(5):353–361, 2014.

Austin Talbot and Yue Ke. Detecting batch heterogeneity via likelihood clustering. *arXiv preprint arXiv:2601.09758*, 2026.

Austin Talbot, Alex Kotlar, Lavanya Rishishwar, and Yue Ke. Classifying copy number variations using state space modeling of targeted sequencing data: A case study in thalassemia. *arXiv preprint arXiv:2504.10338*, 2025.

Austin Talbot, Alex Kotlar, Lavanya Rishishwar, Andrew Conley, Mengyao Zhao, Nachen Yang, Michael Liu, Zhaohui Wang, Sean Polvino, and Yue Ke. Bayescnv: A bayesian hierarchical model for sensitive and specific copy number estimation in cell free dna. *Diagnostics*, 16(2):280, 2026.

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.

Harry Urkowitz. Energy detection of unknown deterministic signals. *Proceedings of the IEEE*, 55(4):523–531, 1967.

Antonio C Wolff, M Elizabeth Hale Hammond, Kimberly H Allison, Brittany E Harvey, Pamela B Mangu, John MS Bartlett, Michael Bilous, Ian O Ellis, Patrick Fitzgibbons, Wedad Hanna, et al. Human epidermal growth factor receptor 2 testing in breast cancer: American society of clinical oncology/college of american pathologists clinical practice guideline focused update. *Archives of pathology & laboratory medicine*, 142(11):1364–1382, 2018.

Appendix A. Derivational Details

A.1. Moment-Matching for the Gamma Approximation

Let $Y = \tau_j^2\,\chi_1^2(\lambda_j)$ where $\chi_1^2(\lambda_j)$ is the noncentral chi-squared distribution with one degree of freedom and non-centrality λ_j. Its first two moments are

$$\mathbb{E}\big[\chi_1^2(\lambda_j)\big] = 1 + \lambda_j, \qquad \mathrm{Var}\big[\chi_1^2(\lambda_j)\big] = 2(1 + 2\lambda_j).$$

Scaling by τ_j^2 gives

$$\mathbb{E}[Y] = \tau_j^2(1 + \lambda_j), \qquad \mathrm{Var}[Y] = 2\tau_j^4(1 + 2\lambda_j).$$

For the Gamma(a, b) distribution (shape a, scale b) we have $\mathbb{E}[Y] = ab$ and $\mathrm{Var}[Y] = ab^2$. Equating:

$$ab = \tau_j^2(1 + \lambda_j), \tag{18}$$

$$ab^2 = 2\tau_j^4(1 + 2\lambda_j). \tag{19}$$

Dividing (19) by (18):

$$b_j = \frac{2\tau_j^2(1 + 2\lambda_j)}{1 + \lambda_j},$$

and substituting back:

$$a_j = \frac{(1 + \lambda_j)^2}{2(1 + 2\lambda_j)}.$$

Because $(1 + \lambda_j)^2 - (1 + 2\lambda_j) = \lambda_j^2 \geq 0$, it follows that $a_j \geq \frac{1}{2}$, with equality if and only if $\lambda_j = 0$ (no process-mismatch bias). Fitting a Gamma by maximum likelihood therefore anchors at $a_j = \frac{1}{2}$ in the unbiased regime and absorbs the non-centrality into an inflated shape when bias is present.

A.2. Fisher Consistency Constant for the MAD

For $X \sim \mathcal{N}(\mu, \sigma^2)$, the median absolute deviation is $\mathrm{MAD}(X) = \sigma\,\Phi^{-1}(3/4)$, where Φ is the standard Normal CDF. To make the MAD a consistent estimator of the standard deviation σ, one multiplies by

$$c_n = \frac{1}{\Phi^{-1}(3/4)} \approx 1.4826,$$

so that $\hat{\sigma} = c_n \cdot \mathrm{MAD}$ satisfies $\mathbb{E}[\hat{\sigma}] = \sigma$ under normality (Huber and Ronchetti, 2009). The resulting estimator retains the 50% breakdown point of the median while providing an asymptotically unbiased scale estimate.

Appendix B. Effect of Poor Process Matching

In this section we show the impact that a lack of process matching can have on the CNV estimates. All of these 32 samples are normal in a 172 amplicon panel targeting 5 genes for CNV prediction. Of these 32 samples, 10 are substantially degraded. We estimate the

CNVs in the genes using 5 samples as normal, ranging from all 5 coming from the good samples, to all 5 coming from the bad samples. We then bootstrap these estimates over different choices of normal and evaluate the posterior means in the remaining 17-22 good samples. The results are plotted in Figure 6.

We can see that two genes are strongly affected by a lack of process matching, MET and ERBB2. However, they are affected differently, with ERBB2 suffering from dramatic biases while MET has high variance estimates. As a result, the assumption that a lCNR of 0 is not valid when samples are not process matched, and our confidence intervals must reflect this.

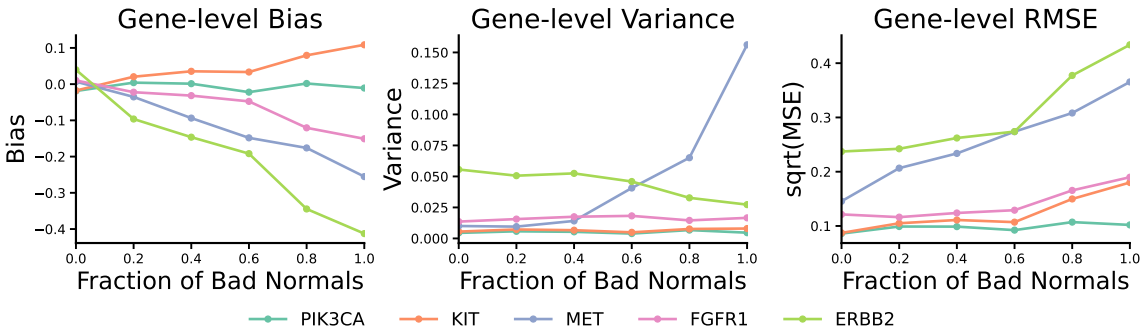

Figure 6: The bias/variance decomposition of CNV estimates as a function of normal matching